Small Area Estimation of Sports Participation at MSOA level using the 2011-12 Active People Survey

Alex Gibson, RAE Consulting Ltd.
&
Paul Hewson, University of Plymouth

August 2014

Contents

1 Research Objectives

1 This project's principal objective is to apply Small Area Estimation methods to Sport England's 2011-12 *Active People Survey 6 (APS6)* in order to generate local area estimates of the number and proportion of adults living in households who:

(a) participate in at least 30 minutes of sport at moderate intensity at least once a week (Sport England's '1×30 indicator'), and

(b) participate in sport and active recreation, at moderate intensity, for at least 30 minutes on at least 12 days out of the last four weeks – equivalent to 30 minutes on three or more days a week (the former National Indictor 8 (NI8) for Local Authorities).

In addition to the overall count and rate of adults meeting the '1×30' and NI8 thresholds, we provide a series of age-sex specific counts and rates; namely for males and females in the seven agebands 16-24, 25-34, 35-49, 50-64, 65-74, 75-84, and 85 and above. The resulting database of local area estimates (presented as a series of Excel spreadsheets in which we provide upper and lower 95% confidence intervals as well as mean estimates) is accompanied by a brief Technical Report (this document).

2 Small Area Estimation (SAE) has been implemented for the 6,791 Middle Layer Super Output Areas (MSOAs) that were defined for disseminating 2011 Census data. (Note that these differ from the 6,781 MSOAs created in 2004 to improve and facilitate the reporting of small area statistics[1]). The 2011 MSOAs nest within current Local Authorities (n=326), and our MSOA-level modelled estimates are also aggregated and presented at LA level.

3 The MSOA-level estimates have been generated using fully Bayesian hierarchical Generalised Linear Mixed Models (GLMM) which link outcome data from the *APS6* with individual-level covariate data drawn from the 2011 Census and MSOA and LA-level data from a variety of supplementary data sources.

4 This Technical Report aims to provide a brief description of the methods and sources used in the study. Thus following a general introduction to the approach adopted (Section 2), this report details the principal methodological and evidential issues that have arisen (Sections 3 and 4) and presents key findings which shed light on the insights that can be gained through small area estimation (Section 5). The report concludes with a guide to how the local area estimates are presented in a series of spreadsheets (Section 6).

[1] See http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/super-output-areas--soas-/index.html for a discussion of the development of a new set of MSOAs for the dissemination of 2011 Census data.

2 Small Area Estimation: An Overview

5 Small Area Estimation (SAE) aims to overcome the problem that whilst many surveys are designed and undertaken at a national level, practitioners often require information about more local areas. Unfortunately, the sample size achieved by most national surveys is usually far too small for direct estimation at the sub-regional level. Despite the scale of the Active People Survey, which obtains a minimum of 500 interviews in most local authorities in England and was designed to enable direct estimates to be made at local authority level, it cannot support direct estimation for geographies below local authority level.

6 Thus although almost all MSOAs are represented in the *APS6* dataset (6,767 of 6,791: 99.6%) there are, on average, only 20.4 respondents per represented MSOA – and the actual number of respondents per MSOA ranges from 1 to 121. The *APS6* is therefore unable to support direct estimation at the MSOA level.

7 Faced with such a scenario, researchers have sought to generate local estimates from national surveys using either indirect standardisation or a multilevel model-based approach. Indirect standardisation, although computationally straight-forward, is problematic because, by simply applying national or sub-national prevalence rates to local populations, it assumes spatial invariance. In other words, whilst it captures 'compositional effects' – how the overall prevalence of the response variable varies from place to place simply because of how the socio-economic composition of populations varies – it cannot capture any 'contextual' effects that may affect its local prevalence.

8 As it seems intuitively plausible that there will be some sort of contextual (area) effect on local rates of sports participation, it is necessary to turn to multilevel model-based approaches. Such approaches interrogate survey data in order to derive models which best describe how a dependent variable responds to individual ***and*** area-level predictor variables drawn from the survey and other sources. Local area estimates are thus calculated by applying the model's parameter estimates to the corresponding covariate values for the local areas. In effect, the goal is to 'pool' evidence from across the wider sample in order to 'enhance' local estimates.[2]

9 An early example of this approach was provided by Twigg, Moon and Jones' work estimating the prevalence of smoking at ward-level on the basis of data drawn from the *Health Survey for England* (*HSfE*).[3] Their initial step was to use *HSfE* data to devise a model of individual smoking behaviour using both individual-level variables derived from the *HSfE* itself (such as sex, age and marital status) and area-level variables drawn from other sources linked to HSfE respondents through

[2] Gelman, A. and Hill, J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, CUP.

[3] Twigg, L., Moon, G. and Jones, K. (2000) "Predicting small-area health related behaviour: a comparison of smoking and drinking indicators", *Social Science and Medicine*, 50: pp1109-20.

place of residence (such as the percentage of private rented households in an area or the percentage of households with two or more cars). The model parameters (and their interactions) were then used to estimate for each ward the proportion of smokers in each combination of age, sex and marital status, and these proportions were then applied to the corresponding census counts to provide an estimate of smoking prevalence. A number of other studies have since adopted similar multilevel approaches.[4]

10 The use of multilevel models which combine individual- and area-level effects represents a significant advance in small area estimation, but individual-level variables must be identically defined in both the survey from which the model was derived and the supplementary datasets from which area-level covariate data are drawn. In practice, this means only *APS6* variables which match, or can be made to match, variables available in the *2011 Census* can be used in the final model. As discussed further in Section 3 below, this does constrain variable selection.

11 A more significant challenge is that it is not easy within the classical ('frequentist') statistical framework to quantify the precision of small area estimates without simplifying assumptions or the use of computationally-intensive bootstrapping techniques. In other words, the calculation of confidence intervals around estimates of proportion of adults meeting the '1×30' and NI8 targets for sports participation in each local area is not straightforward. Recent advances have, however, made feasible an alternative statistical framework which focuses on generating 'posterior' distributions of 'possible' estimates for each small area. Thus, rather than producing a single estimate of, say, the proportion of people meeting the NI8 participation target – around which a theoretically-derived confidence interval is placed – the goal is to define what is termed a 'posterior distribution' of many simulated possible outcomes *given the data being modelled*.

12 Advocates of this *Bayesian approach* to statistical modelling argue that it is more appropriate to focus in this way on conditional ('posterior') distributions of unknown quantities – i.e. conditioning the model on the data – rather than, as in conventional 'frequentist' modelling, on conditioning the model on the basis of the distribution of a test statistic which assumes a range of unseen possibilities for the data. In brief, in the Bayesian approach parameters are thought of as 'random quantities' (rather

[4] Gibson, A., Bailey, T, and Fraser, D. (2004) *Demographic mapping of the 2003 Skills for Life Survey to local areas*. Technical Report for the Department for Education and Skills, December 2004; Heady, P. *et al*. (2003) *Small Area Estimation Project Report*. Model-Based Small Area Estimation Series No.2, ONS Publication; Longhurst, J., Cruddas, M. and Goldring, S. (2005) *Model-based Estimates of Income for Wards, 2001/02: Technical Report*. Published in Model-Based Small Area Estimation Series, ONS Publication; Bajekal, M., et al.. (2004) *Synthetic estimation of healthy lifestyles indicators: Stage 1 report*. National Centre for Social Research. (Available at http://old.iph.ie/files/file/Synthetic_Estimation_Stage_1_Report.pdf.) [Accessed 29/10/2013]; BIS Research Paper 81D, *2011 Skills for Life Survey: Small Area Estimation User Guide* (http://tinyurl.com/pfrlpns) and BIS Research Paper 81C, *2011 Skills for Life Survey: Small Area Estimation Technical Report* (http://tinyurl.com/o99r2wt) – with data available at https://www.gov.uk/government/statistical-data-sets/2011-skills-for-life-survey-small-area-estimation-data [all accessed 29/10/2013].

than fixed constants as in classical statistics) so that the statistical model sought is not just the *likelihood function* $P(y \mid \theta)$ which is the probability density of the data $y = y_1, \ldots, y_n$ 'given' the parameter values $\theta = \theta_1, \ldots, \theta_n$, i.e. $P(y \mid \theta)$, but rather the joint probability distribution for both the data and the parameters, i.e. $P(y, \theta)$.

13 $P(y, \theta)$ is linked to the likelihood function via $P(y, \theta) = P(y \mid \theta) P(\theta)$ where $P(\theta)$ is known as the ***prior*** probability distribution for the parameters because it expresses uncertainty about θ before taking the data into account. It is usually chosen to be 'non-informative'. Bayes' Theorem then allows the ***posterior*** probability distribution for the parameters to be derived given the observed data:

$$P(\theta \mid y) = \frac{P(y \mid \theta) P(\theta)}{P(y)} = \frac{P(y \mid \theta) P(\theta)}{\int_{\theta} P(y \mid \theta) P(\theta) d\theta}$$

14 In other words, the 'posterior' is proportional to 'likelihood' × 'prior' – the denominator being a normalising constant independent of the parameters. All information concerning the parameters (and functions of them, such as predictions) can then be derived for the relevant posterior distribution. Whilst in principle this offers a general and flexible approach to statistical modelling which is capable of handling very complex modelling frameworks, the problem has always lain with the complex integrations required to evaluate the denominator involved in the expression for the posterior distribution. Indeed, whereas many modern 'real-world' implementations are faced with large numbers of random effects (as in the present instance where we have a random effect for each of the 326 LAs and 6,767 MSOAs covered by the survey), in practice all but the very simplest integrations are, to all intents and purposes, mathematically intractable.

15 Over the past decade, however, the development of Markov chain Monte Carlo (McMC) simulation techniques – and computers with sufficient power to carry out the necessarily intensive calculations – has overcome the need for complex numerical integration and made Bayesian modelling of complex situations involving many parameters a practical feasibility.

16 At the heart of the McMC simulation-based approach is the construction of Markov chains with particular conditional probability density functions (i.e. the probability of one parameter given all the other parameters at their currently estimated values and the data) as their equilibrium distribution. We draw a random sample from the conditional probability density function and this draw becomes the new value of that parameter and the simulation continues to iterate. The McMC simulation is run for a long time so that it converges and sample values are collected. If such samples are numerous and the chain has properly converged this provides virtually complete information about the required posterior distribution.

17 The principal difficulty with McMC is that, in addition to ensuring the quality of model fit, we have to ensure that the algorithm behaves correctly. Good practice suggests fitting the model several times using different starting points to ensure that the

simulation converges to the same values (these models can be quite complex and we need to be sure the algorithm doesn't find a local "best" solution at the expense of a global "best" solution). Moreover, in order to avoid potential autocorrelation in the simulated values, it is also necessary to 'thin' the converged simulations by throwing away nine values in ten so that what is left is an independent random sample from the distribution of interest. This is computationally massively intensive, but with multiple McMC runs, we can then compare the different sets of simulated values to check that the simulation has reached a viable settled state.

18 McMC algorithms (such as the Gibbs sampler) have made Bayesian modelling of complex situations involving many parameters a practical feasibility[5]. The small area estimation problem provides just such a situation, and the Bayesian approach, combined with associated McMC techniques, provides a unified and flexible framework within which suitable multilevel models (involving both individual and area level covariates and both fixed and random effects) can be fitted to individual survey data and then used to generate posterior predictive distributions for small area estimates.

Bayesian Mixed-Effects (Multilevel) Models for Survey Data

19 The generality and flexibility of the Bayesian approach means that it can deal with a wide range of problems including, as in the present instance, the specification of complex mixed-effects or multilevel models. Multilevel models are so called because of the hierarchical (or 'multilevel') structure by which the data have been collected and/or within which processes are presumed to operate. Individuals are thus 'nested' within one or more areas and the likelihood that they will, for instance, meet the NI8 and '1×30' targets for sports participation are assumed to be a function of both their individual social-demographic characteristics and aspects of the upper level groups of which they are a part (in this instance, their MSOA and LA populations).

20 Multilevel models are of particular importance in small area estimation because of the way in which information is 'borrowed' about individuals in other areas. If this is done without accounting for correlation induced by similar area characteristics, independence assumptions (formally 'exchangeability assumptions' in a Bayesian setting) are violated. Multilevel models with 'fixed' (individual-level) and 'random' (area-level) effects – hence 'mixed-effects models' – have thus become increasingly popular in the last two decades. Precisely because the approach takes into account all uncertainty associated with unknown model parameters, the advantages of adopting a fully Bayesian approach to small area estimation have long been aired in the literature.[6]

---

[5] Congdon, P. (2001), *Bayesian Statistical Modelling*. Chichester: Wiley; Congdon, P. (2003), *Applied Bayesian Modelling*. Chichester: Wiley.

[6] Moura, F.A.S. and Migon, H.S. (2002). "Bayesian spatial models for small area estimation of proportions", *Statistical Modelling*, 2: pp183-201; Pfeffermann, D. (2002) "Small Area Estimation – New Developments and Directions", *International Statistical Review* 70(1): pp125-143;

21 The key point is not that multilevel models cannot be fitted any other way (some can), but that the Bayesian approach is more straightforward, can deal with a wider range of models and provides more comprehensive information about the estimates generated. That is to say, by adopting a Bayesian approach and thereby modelling the full posterior predictive distribution of estimates (in effect generating a large number of independent estimates of, for instance, the number of adults participating in at least 30 minutes of sport at moderate intensity at least once a week) it is possible to derive empirically both a 'point estimates' of the number of adults hitting the '1×30' and NI8 participation targets in each MSOA (i.e. the means of the posteriors for each MSOA) and 95% 'credible intervals' around those point estimates (the range within which 95% of the posterior estimates lie). This literally defines the range within which we are 95% certain the true value lie.

22 We have thus adopted a Bayesian approach to calculate the number and proportion of adults in local areas who meet the NI8 and '1×30' participation thresholds. As described in the next sections, we have modelled these statistics on the basis of data drawn from the *Active People Survey 6*, *(2011-12)* and applied the derived parameter estimates and their distributions to corresponding covariate values for local areas (as drawn from the *2011 Census* and other supplementary datasets). To accomplish this we have used the public domain and widely used *stan* software[7] – a program for Bayesian analysis of complex statistical models using Markov chain Monte Carlo (McMC) techniques – with data pre-processing and post-processing being carried out using SPSS[8], the R statistical software package[9], and MySQL[10].

Pfeffermann, D. (2013) "New important developments in small area estimation", *Statistical Science* 28(1): pp40-68.

7 Stan Development Team (2013), *Stan: A C++ Library for Probability and Sampling, Version 2.0*. (Available at http://mc-stan.org.) [Accessed on 31/10/2013.]

8 IBM Corp. Released 2012. *IBM SPSS Statistics for Windows, Version 21.0*. Armonk, NY: IBM Corp.

9 R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. (Also see http://www.R-project.org.) [Accessed on 31/10/2013.]

10 Oracle (2013) MySQL 5.5 (Available from http://www.mysql.com/) [Accessed on 3/10/2013.]; David A. James and Saikat DebRoy (2010). *RMySQL: R interface to the MySQL database*. *R package version 0.7-5.*(Available at http://CRAN.R-project.org/package=RMySQL.) [Accessed on 12/10/2012.]

3 APS6 Small Area Estimation: Model Specification & Fitting

23 Our approach to Small Area Estimation is thus based on fitting hierarchical (or 'multilevel') mixed effects models to data drawn from the *APS6* and supplementary sources, and then applying the resulting model parameter distributions to corresponding covariate values for local areas. As noted above, the dependent variables to be modelled are;

(a) whether individuals participate in at least 30 minutes of sport at moderate intensity at least once a week ('1×30'), and

(b) whether individuals participate in sport and active recreation, at moderate intensity, for at least 30 minutes on three or more days a week (NI8).

24 Both dependent variables are binary and in each case we apply a binary logistic regression model with a two-category response y_{ijk} where $y_{ijk}=1$ if individual i within MSOA j and Local Authority k meets the relevant participation criteria, and $y_{ijk}=0$ if they do not, and where;

$$y_{ijk} \sim Bernoulli\left(p_{ijk}\right); \qquad p_{ijk}=\frac{\exp(\eta_{ijk})}{1+\exp(\eta_{ijk})}$$

η_{ijk} is specified as a linear predictor βX comprising individual and upper (MSOA and LA) level variables. In general terms, the mixed effects binary logistic regression thus takes the form;

$$\eta_{ijk}=\beta_{0j}+\beta_{0k}+\beta_0+\beta_1\cdot\text{sex}_j+\beta_2\cdot\text{ageband}_j+\beta_3\cdot\text{GeneralHealth}_j$$
$$+\beta_4\cdot\text{Ethnicity}_j+[\text{other individual-level covariates}]$$

where

$$\beta_{0j}=\gamma_0+\gamma_1\cdot\%\text{GCSE}_j+\gamma_2\cdot\text{Pop-Turbulence}_j+[\text{other MSOA covariates}]+\varepsilon_j$$

$$\beta_{0k}=\gamma_0+\gamma_1\cdot\text{clubspp}_k+[\text{other LA covariates}]+\varepsilon_k$$

$$\varepsilon_j=N(0,\sigma^2)$$

25 These models aim to incorporate MSOAs as an upper level because they provide a suitably local and well-established geographical framework for which an appropriate range of relatively up-to-date area data are available, and LAs in order to align the model with the geography that underpinned the survey's sampling framework as well as to capture any policy-related effects which might (in theory) be expected to affect participation rates.

26 The *APS6* dataset comprises 163,420 individuals, all of whom provide a valid response to the question about participation in at least 30 minutes of sport at moderate intensity at least once a week (henceforth the '1×30' question). There were, however, 355 non-responses to the question about participation in sports and active recreation over the previous four weeks, resulting in a sample of 163,065 respondents upon which to analyse responses to the NI8 question.

27 Individual-level effects are estimated using data drawn from the survey itself, but area-level effects are estimated using data derived from a variety of supplementary sources. These provide a range of relatively up-to-date measures which describe various aspects of local areas (MSOAs) and local authorities – and which in many cases act, to some degree, as proxies for a broader set of social, economic and geographical characteristics that differentiate between places. These upper-level supplementary variables can be utilised because each individual in the *APS6* is explicitly assigned to their local authority, and the vast majority (136,996: 83.8%) have been assigned to a MSOA of residence on the basis of their postcode.

28 The *APS6* offers a wide range of potential individual-level variables but, as discussed further in Section 4 below, these must match data on local areas available in the *2011 Census*. This excluded some potentially important variables (such as whether or not respondents had a long-standing illness and their highest educational qualifications). Table 1 below lists the candidate individual-level variables (and their factors) available to the analysis once this requirement to align with 2011 Census variables had been met.

Table 1 Individual-level variables and factors available for modeling

Variable	Factors
sex (2 factors)	Male; Female
ageband (7)	16-24; 25-34; 35-49; 50-64; 65-74; 75-84; 85+
genhealth (5)	Very good; Good; Fair; Bad; Very bad
ethnic (5)	White; Mixed; Asian/Asian British; Black/Black British; Other
econact (8)	Employed/Self-employed (FT); Employed/Self-employed (PT); Unemployed; Retired; Student (EA inactive); Looking after family/home; Long-term sick/disabled; Other EA Inactive
hhtype (4)	Single person; Single parent; Adult-only family; Family with children
nssec (9)	Higher Managerial & Professional; Lower Managerial & Professional; Intermediate; Small employers & own account workers; Lower supervisory & technical; Semi-routine; Routine; Never worked or LT Unemployed; Not classified
religion (8)	Christian; Buddhist; Hindu; Jewish; Muslim; Sikh; Other; No religion
tenure (5)	Owned outright; Owner with mortgage; Rented from council; Rented from Housing Association; Privately rented or living rent free
cars (3)	No cars; 1 car; 2+ cars

29 The MSOA- and LA-level covariate data available to the analysis are listed in Table 2 and Table 3 below. An intensive search was undertaken (given the time available) for likely MSOA-level variables, and we have made use of data presented using both pre- and post-2011 MSOAs. Data only available for pre-2011 MSOAs have been attributed to current (post-2011) MSOAs using a population-weighted algorithm as discussed in the appendix (Section 7) .

Table 2 MSOA-level covariate data available for modeling†

From 2011 Census Data (NOMIS)	
densitypph	Density (persons per hectare) (QS102EW)
L4plusQuals	% adults with L4 qualifications or higher (QS501EW)
L3plusQuals	% adults with L3 qualifications or higher (QS501EW)
From Public Health England Local Health Indicators:	
goodchild	% 5 year olds with good development
childpov	% children in households in poverty
hhpov	% pop in households receiving means-tested benefits
GCSE	% students achieving 5+ GCSEs (incl Maths & English)
LTunemp	% working age pop unemployed for 12 months or more
From ONS MSOA Population Turnover Rates, mid-2009 to mid-2010	
TotTurbulence	Total GP Population Turbulence (in+out/pop)
NetFlow	Net GP Population Turbulence (in-out/pop)
TT1564	Age15-64 Total GP Population Turbulence (in+out/pop)
NF1564	Age15-64 Net GP Population Turbulence (in-out/pop)
From ONS Neighbourhood Statistics Modelled Healthy Lifestyle Behaviours	
smoking	Modelled estimate of % smokers
BingeDrink	Modelled estimate of % binge drinkers
obese	Modelled estimate of % obese
Veg5aday	Modelled estimate of % achieving 5-a-day veg and fruit
From English Indices of Deprivation 2010: Scores and Underlying Indicators	
IMDscore	IMD2010 - Pre-2011 LSOA values assigned to 2011 MSOAs
IMDincome	IMD2010 Income Domain - assigned to 2011 MSOAs
IMDemployment	IMD2010 Employment Domain - assigned to 2011 MSOAs
IMDhealth	IMD2010 Health Domain - assigned to 2011 MSOAs
IMDeducation	IMD2010 Education Domain - assigned to 2011 MSOAs
IMDservices	IMD2010 Barriers to Services Domain - assigned to 2011 MSOAs
IMDcrime	IMD2010 Crime Domain - assigned to 2011 MSOAs
IMDenvironment	IMD2010 Living Environment Domain - assigned to 2011 MSOAs
Not_HE	% people under 21 not entering Higher Education
Dist2FdShp	Average distance (km) to Food Shop (OA centroid to shop)
Dist2PriSch	Average distance (km) to Primary School (OA centroid to school)
Acute	Age standardised rate of emergency admissions to hospital
LowIncome	% income deprived individuals
Unemploy	% employment deprived individuals

† See Appendix 1 for further details concerning the construction and provenance of these variables.

30 Only one LA-level covariate dataset has been included, in part because of a lack of potentially explanatory variables at LA level, but also because, as it transpired, our models suggest there is little LA-level variance to be explained once individual and MSOA-level effects are included in the model. Certainly, as discussed below, potential LA-level covariate effects do not contribute significantly to any of our candidate models and, in the end, we have not included any LA-level covariates in the final models. Nor have we retained a random effect for each local authority in

the final model. This would have provided a LA-level intercept value which allows the estimates for all MSOAs in a local authority to vary in response to variations in participation at an explicitly local authority level. Exploratory modelling showed that, once we have accounted for individual- and MSOA-level effects, LA-level random effects are negligible – ranging from -4.63×10^{-6} to 3.55×10^{-3} in the '1×30' model and from -7.65×10^{-2} to 6.63×10^{-2} in the NI8 model. Including these LA random effects would have incurred very significant additional computational complexity for no appreciable impact on the estimates.

Table 3 LA-level variables available for modeling†

From Sport England Clubmark Accredited Clubs (as of 14/10/2013)	
Clubspp	Clubmark Clubs per 10,000 persons
Clubspp1664	Clubmark Clubs per 10,000 persons aged 16-64
Clubspp16plus	Clubmark Clubs per 10,000 persons aged 16-64
clubsphec	Clubmark Clubs per 1,000 hectares

† See Appendix 1 for further details concerning the construction and provenance of these variables.

Variable Selection and Model Fitting

The '1×30' Indicator model

31 Not all candidate variables were eventually used in the models. Parameter selection was undertaken using systematic selection procedures based on minimising the Akaike Information Criterion (AIC). In view of the time it takes to fit each multilevel mixed effects model (hours rather than minutes), this process was separated into three distinct stages, each of which focussed on that subset of *APS6* respondents for whom there is no missing data. (Almost the full dataset was used when the selected models were formally fitted using McMC methods.)

32 The first stage involved fitting individual-level models using SPSS binary logistic regression (a process which included a systematic search for possible interaction effects) in a progressive specification of models – each of which would require additional 2011 Census tables when generating the microsimulated population to which the model's individual-level parameter estimates would eventually be applied. The issue here is that without 2011 Census microdata (a 3-5% sample of individual-level census returns due to be published in late-2013/early-2014) the introduction of additional tables at the microsimulation stage inevitably introduces a degree of unquantifiable uncertainty. As discussed in Section 4 below, by minimising the number of tables needed to microsimulate the required full 'joint distribution' (i.e. the detailed socio-demographic composition) of each area, we also minimise the potential uncertainty. There is thus a trade-off between improving model fit with additional variables and increasing uncertainty at the microsimulation stage by using additional census tables.

33 To illustrate, our base model for the '1×30' response variable (which incorporates only ageband7, sex2 and genhealth3) returns a Cox & Snell adjusted r^2 value of 0.128 and an AIC value of 38,576. The model is not particularly good (nor does it account for characteristics in which we are intrinsically interested – such as ethnicity or social class), but at least the parameter estimates can be applied to ***known*** covariate data because Census Table DC3302EW gives counts of the number of people in each MSOA in all (7*2*3=42) cells of the required joint distribution.

34 Systematically increasing the number of variables (and appropriate interaction effects) to minimise the AIC results in a demonstrably better model fit, but it also introduces uncertainty at the microsimulation stage. Thus our final 'full' model for the '1×30' response variable incorporates 7 individual-level variables (ageband7, sex2, genhealth5, nssec9, tenure5, ethnic5 and cars3), along with 5 interaction effects (ageband7*sex2, ageband7*genhealth5, sex2*nssec9, ethnic5*sex2 and ethnic5*cars3). This returns a much improved Cox & Snell adjusted r^2 value of 0.166 and a significantly smaller AIC value of 37,305. The model is much better, but the penalty is unquantifiable additional uncertainty due to the fact that we must now reconcile 6 separate census tables in our microsimulation of MSOA populations (namely DC3302EW, DC3601EW, DC6303EWR, DC6206EW, DC1104EW and DC4203EW).

35 Until 2011 census microdata are available to constrain the microsimulation of MSOA populations we have to accept this unquantifiable uncertainty, although it will be of limited significance. It should nevertheless be recognised, as discussed further in Section 4 below, that microsimulation may mis-specifiy the detailed socio-economic composition of MSOAs with, in particular, unusual patterns of tenure and/or car ownership relative to age, sex, general health and social class (NS-SEC). Whilst not optimal, it seems to us unlikely that this will materially affect the final estimates.

36 Having defined the full set of individual-level variables to be included in the final model, the final step in this initial stage was to explore whether any of the variables could be collapsed without unduly affecting model fit. This was primarily to minimise the risk that the final mixed effects hierarchical models would fail to converge satisfactorily – a risk that increases with model complexity (including the total number of factors). To that end we were able to reduce (a) ethnicity from 5 to 4 factors (conflating the 'white' and 'mixed' categories); (b) tenure from 5 to 3 factors (conflating the 'owned outright' and 'owned with a mortgage' categories into a single 'owned' category, and the 'rented from council' and 'other social rented' categories into a single 'socially rented' category); and (c) NS-SEC from 9 to 8 factors (conflating 'Higher Managerial and Professional Occupations' and 'Lower Managerial and Professional Occupations' into a single category).

37 The resulting 'optimal' individual-level '1×30' model thus comprises seven main effects (sex2, ageband7, genhealth5, nssec8, tenure3, ethnic4 and cars3) contributing 32 factors, and five interaction effects contributing a further 46 factors

(ageband7*sex2, ageband7*genhealth5, sex2*nssec8, ethnic4*sex2 & ethnic4*cars3). With a Cox & Snell adjusted r^2 value of 0.166, the model coefficients and odds ratios for the main effects are as given in Table 4 below.

Table 4 Individual-level Parameter Estimates: the '1×30' model

	Coefficients		
Factor [Ref. Group]	**B**	**SE**	**Odd Ratios**
Constant	0.289	0.100	
Sex [Female]			
Male	1.098	0.104	1.40
Ageband [16-24]			
25-34	-0.171	0.109	0.93
35-49	-0.315	0.098	0.86
50-64	-0.914	0.099	0.61
65-74	-1.092	0.108	0.54
75-84	-1.417	0.130	0.43
85+	-2.371	0.276	0.19
General Health [Very Good]			
Good	-0.162	0.092	0.93
Fair	-0.542	0.128	0.76
Bad	-0.423	0.252	0.82
Very Bad	-0.073	0.579	0.97
NS-SEC† [NS-SEC 1]			
NS-SEC 2	-0.202	0.049	0.91
NS-SEC 3	-0.163	0.066	0.93
NS-SEC 4	-0.451	0.071	0.80
NS-SEC 5	-0.459	0.050	0.80
NS-SEC 6	-0.603	0.075	0.74
NS-SEC 7	-0.434	0.095	0.81
NS-SEC 8	-0.101	0.076	0.96
Tenure [Owned]			
Social Rented	-0.437	0.045	0.81
Other rented and other	-0.223	0.041	0.90
Ethnicity [White or mixed]			
Black	-0.283	0.165	0.88
Asian	-0.237	0.179	0.90
Other	0.010	0.301	1.00
Cars [No cars]			
1 car	0.413	0.041	1.17
2+ cars	0.602	0.044	1.24

38 The odds ratios are much as one might expect, though they are mediated by interaction effects that we have not included in Table 4. For instance males are, on average, about 1.4 times more likely to fulfil the '1×30' criteria than females; whilst older people are progressively much less likely than younger people to report that they engage in moderate intensity sport for at least 30 minutes a week. Occupants of rented accommodation, and particularly of socially rented accommodation, are less likely than owner occupiers to meet the '1×30' threshold; as are Black and Asian respondents relative to those of white or mixed ethnicity. People with access to a car, and particularly those with access to 2+ cars, are more likely to have claimed to meet the '1×30' threshold than those without access to a car; as are

those in the managerial and professional occupations reference category relative to the other NS-SEC occupational categories.

39 It is worth emphasising, however, that these relationships are not being presented as *causal* or *explanatory*. Some of these factors may be (and have been discussed as such elsewhere[11]), whilst other factors (such as access to a car) most likely stand as proxies for a broader set of associated socio-economic characteristics, but in the context of Small Area Estimation we are concerned only with *prediction* – with identifying relationships between individual-level factors and whether or not individuals claim to have met the '1×30' or NI8 participation criteria. These relationships, at the individual, MSOA and LA level, are then used to estimate local participation rates, without making any claim as to the specific factors which actually influence the behaviour of individuals.

40 Having identified the variables (and interaction effects) to be entered at the individual level of our proposed hierarchical mixed effects model, the second stage of the model specification process was to explore which upper-level candidate variables best contribute to the proposed model's capacity to predict whether or not individuals in the *APS6* dataset meet the '1×30' participation criteria. To that end we initially used the R Statistics *glmer* function (part of the *lme4* library[12]) to fit, in turn, each of the candidate MSOA-level variables reported in Table 2 above. As reported in Table 5 below, which is sorted by increasing AIC values, the majority of the candidate variables are significant and would act to shift local estimates in largely predictable directions.

41 It is not coincidental that the most important MSOA-level variables are related to education. It proved difficult to match how individuals' qualifications were recorded in the *APS6* with 2011 census data on highest qualifications. This potentially important predictor of propensity to engage in sport is thus missing from the individual-level model. It is clearly being picked up at a higher level in the mixed effects model.

42 In theory one can now parallel what was done at an individual level and explore what combination of MSOA-level covariates maximises model fit, but this is a time-consuming exercise. Whereas it might take a matter of minutes to test the impact of adding or removing a candidate variable to a simple binary logistic regression model, fitting a multilevel mixed effects logistic model to the *APS6* dataset using *glmer* can take much longer – even when using the subset of 35,203 individuals

[11] Sport England, *Understanding variations in sports participation. Technical Report* (22/4/2010) (Available at http://tinyurl.com/okxh44v) [Downloaded 12/10/2013].; Foster, C. et al., *Understanding Participation in Sport – a Systematic Review. A study on behalf of Sport England by the University of Oxford British Heart Foundation*. (Sport England, March 2005) (Available at http://www.sportengland.org/media/39119/understanding-participation-in-sport-2005.pdf) [Downloaded 12/10/2013]..

[12] Douglas Bates et al., *Linear mixed-effects models using Eigen and S4* (the 'lme4' package), 25 October, 2013. (Available at http:// http://cran.r-project.org/web/packages/lme4/lme4.pdf) [Accessed 30/10/2013.]

(in 6,272 MSOAs) for whom there is no missing data. Indeed, it is not uncommon to run a complex model for up to 8 hours, only to find that it has failed to converge and the parameter estimates are of uncertain value!

Table 5 '1×30' model fit introducing candidate MSOA-level covariate data

		Upper-level coefficient		
Covariate name	Model Fit (AIC)	Estimate	S.E.	Sig.
L4plusQuals	37,195	0.0142	0.0013	<0.0001
L3plusQuals	37,204	0.0131	0.0012	<0.0001
Not_HE	37,205	-0.0087	0.0008	<0.0001
obese	37,219	-0.0316	0.0031	<0.0001
IMDeducation	37,227	-0.0395	0.0041	<0.0001
Veg5aday	37,233	0.0204	0.0022	<0.0001
GCSE	37,236	0.0095	0.001	<0.0001
smoking	37,261	-0.0146	0.0019	<0.0001
hhpov	37,291	-0.0093	0.0017	<0.0001
IMDincome	37,291	-0.9251	0.1738	<0.0001
LowIncome	37,291	-0.0092	0.0017	<0.0001
goodchild	37,293	0.0058	0.0011	<0.0001
IMDscore	37,293	-0.0060	0.0012	<0.0001
IMDemployment	37,295	-1.3612	0.276	<0.0001
Unemploy	37,295	-0.0135	0.0027	<0.0001
childpov	37,300	-0.0051	0.0011	<0.0001
IMDhealth	37,300	-0.0811	0.018	<0.0001
LTunemp	37,308	-0.1136	0.032	0.0004
BingeDrink	37,311	-0.0084	0.0029	0.0034
IMDcrime	37,311	-0.0576	0.0202	0.0044
IMDservices	37,314	0.0033	0.0014	0.0190
densitypph	37,315	0.0010	0.0005	0.0351
TT1564	37,315	0.0078	0.0036	0.0304
Acute	37,315	-0.1090	0.0496	0.0280
TotTurbulence	37,317	0.0072	0.0040	0.0700
NetFlow	37,317	0.0274	0.0183	0.1343
NF1564	37,318	0.0082	0.0170	0.6317
IMDenvironment	37,320	-0.0001	0.0001	0.9324
Dist2FdShp	37,320	-0.0015	0.0102	0.8857
Dist2PriSch	37,320	0.0132	0.0282	0.6390

43 Our solution reflects the fact that Small Area Estimation is not concerned with identifying specific factors which may account for differences in participation in sports[13] – but only in developing models which best predict whether or not individuals in the *APS6* dataset meet the '1×30' or NI8 participation thresholds. To that end, the information available on MSOAs has been summarised using Principal Components Analysis (PCA). On this basis a set of four factors were derived which, taken together, account for 75% of the variability of 19 of the

[13] The Sports England report on variations between local areas (LAs and MSOAs?)

variables listed in Table 5 above. (We excluded the overall IMD score and the seven IMD domain scores to avoid possible problems of circularity – these scores were themselves built using factor reduction techniques utilising some of the other variables listed in Table 5. We also excluded the *Dist2FdShp*, *LowIncome* & *Unemploy* variables as these were very closely correlated with other variables).

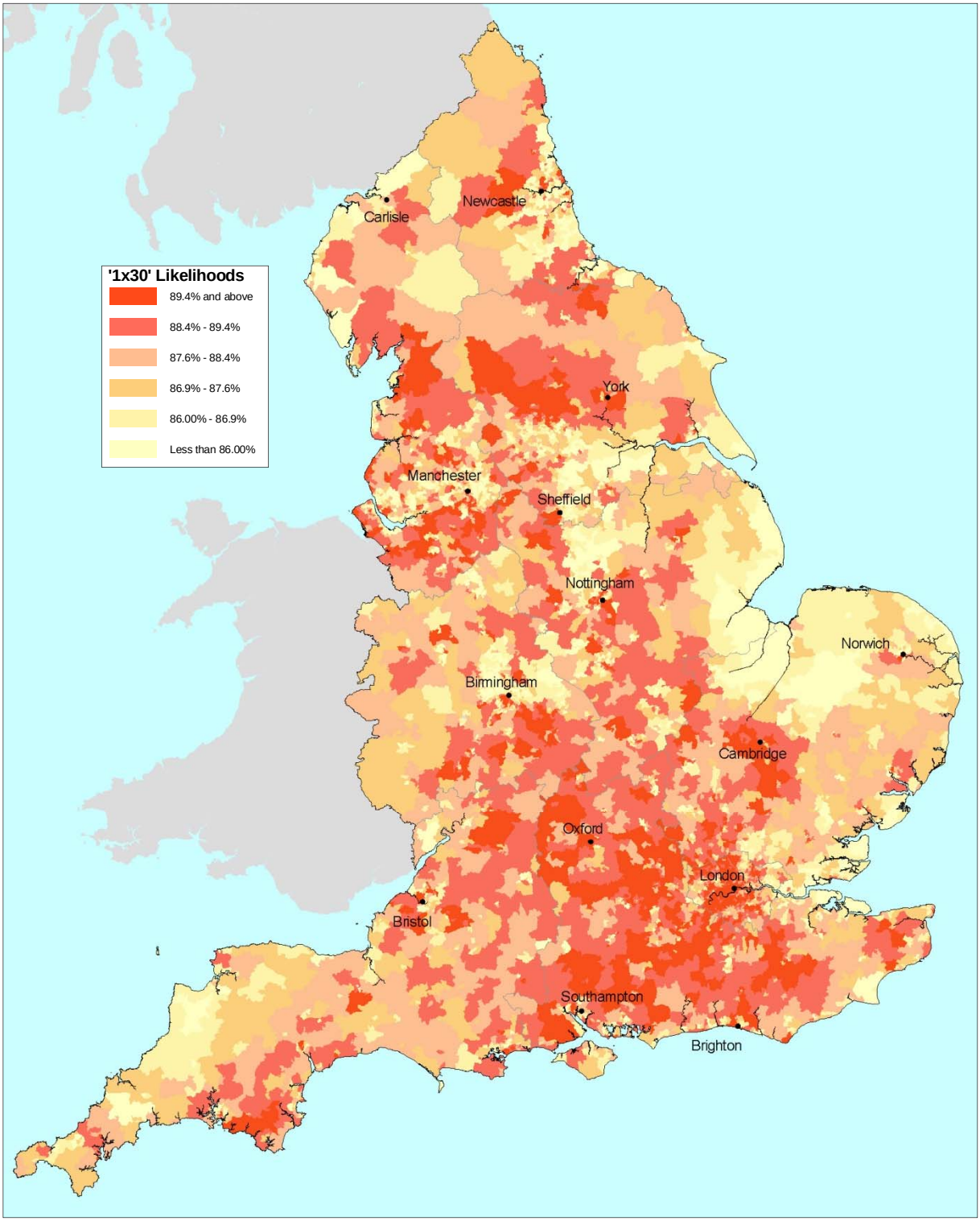
44 Taken together the four factors can be thought of as offering the best means of discriminating between MSOAs in a manner which captures the diversity of information we have, and the resulting model returns an AIC value of 37,196 – suggesting a model fit which virtually matches that achieved by the best single-factor model listed in Table 5 above (i.e. that using *L4plusQuals*).

45 Using the four PCA factors at the MSOA level serves to significantly shift our estimates of the likelihood that particular types of people in different areas will meet the '1×30' participation threshold. For instance, the likelihood that a young (25-34) professional man of white ethnicity in very good health and who owns his own home and has access to 2 or more cars will meet the '1×30' participation threshold varies from 83.3% in the MSOA 'least conducive' to participation (Wyre Forest 009) to 93.0% in the 'most conducive' MSOA (Oxford 003). The manner in which these PCA factors combine to shift likelihoods across the country as a whole, illustrated in Figure 1 below, shows a distinct geographical pattern, and we are satisfied that using PCA factors in this way provides an effective method of maximising our model's predictive strength at the local level.

46 The third and final stage in model specification concerns the introduction of LA-level effects. With few obvious LA-level datasets containing variables which might plausibly explain variations in participation at an explicitly local authority level, we were only able to consider the number of Sport England's Clubmark accredited clubs as a potential measure of LA-level variations in the provision of sporting facilities[14]. As described in Table 6 below, neither the number of clubs *per capita* nor the number per person aged 11-64 proved to be significant and no case can be made for their inclusion in the final model.

[14] We attempted to extract meaningful information from data on National Lottery grant awards as it has been suggested that such may provide a proxy measure of the long-term commitment of local authorities to 'investing in the frameworks, people and programmes that support sport' (Sport England, *Understanding variations in sports participation between local authorities* (August, 2010) (See http://www.sportengland.org/media/40389/findings-from-local-variation-modelling-summary.pdf) [Accessed 31/10/2013]). Unfortunately it ultimately proved impossible to distinguish between grants made to local organisations and those made to national organisations but with headquarters in a particular local authority.

Figure 1 Modelled MSOA-level variation of likelihood '1×30' participation



This maps the modelled likelihood in our 2-level (individual & MSOA) mixed effects model that a young (25-34) professional man of white ethnicity in very good health and who owns his own home and has access to 2+ cars will meet the '1×30' participation threshold. It illustrates the degree to which local estimates are shifted in response to variations in the four PCA-factors - which together capture 75% of the ways in which MSOAs differ from one another. The same relative effect would apply to all person types included in the model.

Table 6 '1×30' model fit on introducing candidate LA-level covariate data

	Model Fit		LA-level coefficient		
Covariate name	(AIC)	Converged?	Estimate	S.E.	Sig.
<base model>	37,196	Yes	-	-	-
LA clubs per person	37,195	Yes	-0.01026	0.014936	0.49207
LA clubs per person 11-64	37,196	Yes	-0.00592	0.010029	0.55470

47 In developing appropriate mixed effects models it became apparent that, once MSOA-level random effects are alongside individual and MSOA-level fixed effects, there remains little additional variance to explain at the LA-level. In fact, we doubt whether incorporating any LA-level covariate data would serve to improve the predictive power of our models. Indeed, the negligible size of LA-level random effects meant these too could be excluded from our final model[15]. They would have had no meaningful impact on MSOA-level estimates but would have resulted in a significant additional computational burden on McMC.

48 The final model for the '1×30' indicator thus comprises:

(a) seven main fixed effects at the individual level; namely ageband (7 factors), sex (2), general health (5), NS-SEC (8), ethnicity (4), tenure (3), and car ownership (3),

(b) five interaction fixed effects at the individual level; namely ageband7 BY sex, ageband BY general health, sex BY NS-SEC, ethnicity BY sex, and ethnicity BY car ownership, plus

(c) the four PCA-derived factors as MSOA-level fixed effects

(d) a random effect for each MSOA.

49 This '1×30' model was formally fitted to the 84% of the full *APS6* dataset using *stan* to obtain McMC simulated values for the posterior distributions of all parameters, given the data. The data to which the model was fitted were those with a simple 'missingness' pattern – the impact of complex combinations of missing data being to grossly slow down the McMC simulation which, in the time available, was not an option. Analysis demonstrated that the 84% sample was entirely representative of the *APS6* dataset as a whole. Following standard practice, the posterior distributions were 'thinned' by a factor of 10 and visual checks of trace plots, along with the formal application of Gelman and Rubin's diagnostic tests,[16] confirmed that all posteriors had converged to a steady state.

[15] Note that whilst it is not possible to improve model fit by including LA-level covariates or random effects, this does not mean that more nuanced 'explanatory' models could not be developed to capture, for instance, the impact of policy-related variables at local authority level. Such is not, however, our purpose here.

[16] Gelman, A and Rubin, DB (1992) "Inference from iterative simulation using multiple sequences", *Statistical Science*, 7, pp457-511; Brooks, SP. and Gelman, A. (1997) "General methods for monitoring convergence of iterative simulations", *Journal of Computational and Graphical Statistics*, 7, pp434-455.

50 These posteriors can be summarised in terms of their means and 95% credible intervals – literally the range within which 95% of the simulated values lie. These values are given in Table 7, whilst the posterior distributions themselves are illustrated graphically in Figure 2.

51 A large positive posterior mean attached to a particular factor, for instance being male, indicates that such individuals are more likely to meet the '1×30' participation threshold than people in the reference group (in this case females). Conversely, a large negative posterior mean attached to a particular factor, for instance people aged 85 and above, indicates that such people are less likely to meet the '1×30' participation threshold than people in the reference group (in this case is people aged 16-24).

52 Each factor's credible interval, meanwhile, is indicative of the level of model uncertainty around that factor's posterior mean. Where the number of people in the survey is small as, for instance, with people reporting very bad general health (just 0.74% of the overall *APS6* sample), there is probably too little information to allow for a precise estimate of the corresponding posterior mean. Interaction effects tend to have wide credible intervals for this reason. There are, for instance, only 47 people aged 25-34 who report very bad general health and the credible interval for that interaction effect is correspondingly wide. This is, of course, of relatively little practical significance as in most MSOAs very few people will fall into these cohorts. Nevertheless, it must be recognised that it is possible for a local area to have a relatively large concentration of a cohort which, nationally, is relatively uncommon. In such cases model uncertainty is taken through into the credible interval surrounding the local estimate of the number of people meeting the '1×30' participation threshold.

53 Relatively wide credible intervals can also arise if there is significant diversity of outcome relative to a given factor. Such may partly explain, for instance, the large credible interval for people of 'other' ethnicity. The very slight negative coefficient for this group is swamped by degree of uncertainty surrounding that coefficient – partly because there are relatively few individuals in the *APS6* who fall into this group, but also doubtless because it is something of a 'catchall' category likely to capture cohorts with very different behaviours.

54 Using these posterior values, alongside appropriate MSOA random effects, we estimate the likelihood that each person type in each MSOA will meet the '1×30' threshold. It then becomes a simple (albeit computationally demanding) process of aggregation – first by applying those likelihood estimates to the microsimulated counts of each person type in each MSOA, and then by summing and summarising to the level required. We have, as detailed in Section 6 below, provided Sport England with counts and proportions (with associated confidence estimates) at MSOA and LA level for each of 12 age-sex cohorts, as well as for adults as a whole.

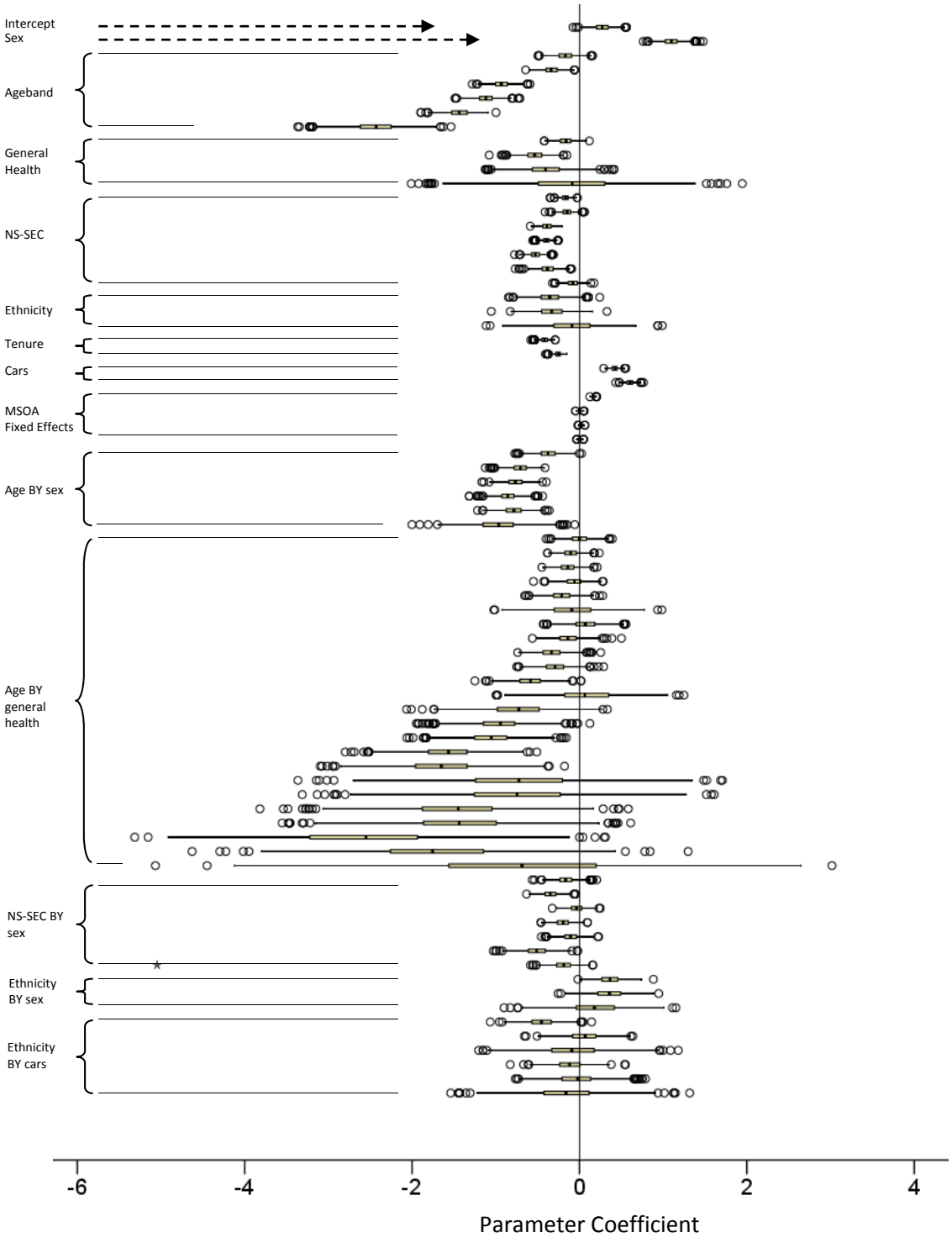
Table 7 '1×30' Model fixed effect posterior means and 2.5 & 97.5% percentiles

Factor [Ref. Group]	mean	2.5%	97.5%
Constant	0.273	0.078	0.466
Ageband [16-24]			
25-34	-0.167	-0.394	0.061
35-49	-0.336	-0.543	-0.146
50-64	-0.932	-1.124	-0.734
65-74	-1.117	-1.318	-0.904
75-84	-1.436	-1.687	-1.183
85+	-2.432	-3.001	-1.881
Sex [Female]			
Male	1.101	0.900	1.315
General Health [Very Good]			
Good	-0.165	-0.346	0.006
Fair	-0.534	-0.784	-0.277
Bad	-0.397	-0.848	0.109
Very Bad	-0.096	-1.281	0.972
NS-SEC† [NS-SEC 1]			
NS-SEC 2	-0.168	-0.259	-0.080
NS-SEC 3	-0.149	-0.274	-0.018
NS-SEC 4	-0.387	-0.525	-0.250
NS-SEC 5	-0.396	-0.487	-0.303
NS-SEC 6	-0.525	-0.668	-0.391
NS-SEC 7	-0.381	-0.570	-0.188
NS-SEC 8	-0.079	-0.233	0.071
Ethnicity [White or mixed]			
Black	-0.354	-0.672	-0.030
Asian	-0.332	-0.667	-0.006
Other	-0.094	-0.705	0.490
Tenure [Owned]			
Social Rented	-0.418	-0.513	-0.335
Other rented & other	-0.256	-0.341	-0.172
Cars [No cars]			
1 car	0.423	0.342	0.508
2+ cars	0.603	0.505	0.696
MSOA Fixed Effects			
PCA Factor 1	0.163	0.136	0.190
PCA Factor 1	0.005	-0.024	0.037
PCA Factor 1	0.022	-0.003	0.048
PCA Factor 1	0.005	-0.022	0.033
Ageband:sex			
25-34:male	-0.375	-0.617	-0.141
35-49:male	-0.707	-0.918	-0.499
50-64:male	-0.767	-0.985	-0.558
65-74:male	-0.856	-1.108	-0.637
75-84:male	-0.785	-1.057	-0.521
85+:male	-0.966	-1.459	-0.414

Factor [Ref. Group]	mean	2.5%	97.5%
Ageband:general health			
25-34:good health	-0.001	-0.254	0.224
35-49:good health	-0.104	-0.300	0.103
50-64:good health	-0.139	-0.337	0.078
65-74:good health	-0.063	-0.296	0.164
75-84:good health	-0.211	-0.483	0.068
85+:good health	-0.091	-0.734	0.521
25-34:fair health	0.076	-0.245	0.431
35-49:fair health	-0.136	-0.419	0.149
50-64:fair health	-0.336	-0.625	-0.059
65-74:fair health	-0.292	-0.609	0.007
75-84:fair health	-0.590	-0.946	-0.252
85+:fair health	0.081	-0.658	0.812
25-34:bad health	-0.737	-1.499	-0.052
35-49:bad health	-0.964	-1.597	-0.392
50-64:bad health	-1.064	-1.691	-0.519
65-74:bad health	-1.579	-2.312	-0.887
75-84:bad health	-1.661	-2.607	-0.800
85+:bad health	-0.712	-2.274	0.963
25-34:v. bad health	-0.747	-2.269	0.773
35-49:v. bad health	-1.455	-2.790	-0.130
50-64:v. bad health	-1.434	-2.703	-0.157
65-74:v. bad health	-2.576	-4.350	-0.856
75-84:v. bad health	-1.741	-3.387	-0.096
85+:v. bad health	-0.705	-3.118	1.652
NS-SEC:sex			
NS-SEC 2:male	-0.164	-0.375	0.030
NS-SEC 3:male	-0.345	-0.522	-0.168
NS-SEC 4:male	-0.034	-0.212	0.140
NS-SEC 5:male	-0.198	-0.371	-0.023
NS-SEC 6:male	-0.105	-0.305	0.097
NS-SEC 7:male	-0.509	-0.821	-0.206
NS-SEC 8:male	-0.189	-0.435	0.046
Ethnicity:sex Ethnicity:sex			
Black:male	0.363	0.096	0.627
Asian:male	0.355	-0.039	0.731
Other:male	0.181	-0.508	0.797
Ethnicity:Access to cars			
Black:1 car	-0.454	-0.802	-0.122
Asian:1 car	0.061	-0.340	0.467
Other:1 car	-0.083	-0.804	0.674
Black:2+cars	-0.118	-0.466	0.227
Asian:2+cars	-0.032	-0.523	0.505
Other:2+cars	-0.165	-1.055	0.673

† Where NS-SEC 1 = 'Managerial & Professional'; NS-SEC 2 = 'Intermediate'; NS-SEC 3 = 'Small employers & own account workers'; NS-SEC 4 = 'Lower supervisory and technical'; NS-SEC 5 = 'Semi-routine'; NS-SEC 6 = 'Routines'; NS-SEC 7 = 'Never worked or LT Unemployed'; and NS-SEC 8 = 'Not classified (FT student, not stated or poorly described)'

Figure 2 '1×30' Model fixed effect parameter estimates



Modelling the NI8 Indicator

55 The process was then be repeated with respect to the NI8 indicator, although not *de novo*. Microsimulating local populations to match the individual-level factors used in a mixed effects model is extremely time-consuming, and a judgement

needs to be made as to whether any improvement that might be made to the NI8 model using a different set of factors warrants the time it takes to generate the required new microsimulation. Our goal here, in other words, must be pragmatic. Does one get a sufficiently robust set of local estimates if we apply the same model structure to the same microsimulated population as has already been undertaken with respect to the '1×30' indicator?

56 In terms of individual-level parameters, as Table 8 below indicates, no advantage is to be gained from either removing variables from the '1×30' model or adding either of the two remaining candidate variables listed on Table 1 above. Nor is model fit improved by replacing the NS-SEC variable by economic activity, which is the only obvious substitution.

Table 8 Adapting the '1×30' model to fit NI8 data

Model	n	df	AIC	Difference
Base model ('1×30')	31,830	71	29,839	-
remove sex2	31,830	54	29,988	149
remove ageband7	31,830	35	30,832	994
remove genhealth8	31,830	43	30,644	805
remove nssec8	31,830	57	29,987	148
remove ethnic4	31,830	59	29,879	40
remove tenure3	31,830	61	29,900	61
remove cars3	31,830	63	29,874	36
Base model ('1×30')	**15,903**	**71**	**14,854**	
add religion8	15,903	78	14,858	4
add econact8	15,903	78	14,859	5
Base model ('1×30')	**31,680**	**71**	**29,749**	
econact8 instead of nssec8	31,680	64	29,891	142
ditto + interaction effects	31,680	71	29,891	142

57 The NI8 can thus appropriately use the same set of individual-level variables as the '1×30' model. Turning to upper-level effects the question, once again, is whether there is any substantial advantage to be gained by using something other than the four factors generated using Principal Components Analysis as described in paragraph 43 above. As illustrated by Table 9 below, a slightly better model fit could be obtained using L3plusQuals (i.e. the % adults with L3 qualifications or higher taken from 2001 Census table QS501EW), but this would narrow the discriminatory focus onto education and we would thus lose the more broadly-based perspective offered by the four PCA factors.

58 Turning finally to LA-level effects, the predictive power of the NI8 model, like the '1×30' model, is not improved by including covariate data based on the number of Clubmark accredited facilities in each local authority. Nor would LA-level random

effects (which range from -7.65x10^{-2} to 6.63x10^{-2}) have anything other than a negligible impact on predictions.

Table 9 NI8 model fit introducing candidate MSOA-level covariate data

		Upper-level coefficient		
Covariate name	Model Fit (AIC)	Estimate	S.E.	Sig.
Base Model ('1×30')	29,795			
> PCA Factor 1		0.1160	0.0163	<0.0001
> PCA Factor 2		-0.0252	0.0181	0.1638
> PCA Factor 3		0.0331	0.0152	0.0296
> PCA Factor 4		0.0009	0.0162	0.9567

Covariate name	Model Fit (AIC)	Estimate	S.E.	Sig.
L4plusQuals	29,787	0.0109	0.0014	<0.0001
L3plusQuals	29,791	0.0100	0.0014	<0.0001
Veg5aday	29,795	0.0176	0.0025	<0.0001
Obese	29,796	-0.0250	0.0036	<0.0001
Not_HE	29,798	-0.0063	0.0009	<0.0001
Smoking	29,806	-0.0136	0.0022	<0.0001
IMDeducation	29,807	-0.0294	0.0049	<0.0001
GCSE	29,818	0.0061	0.0012	<0.0001
LowIncome	29,830	-0.0078	0.0020	0.0001
Hhpov	29,830	-0.0078	0.0020	0.0001
IMDincome	29,830	-0.7812	0.2026	0.0001
Childpov	29,832	-0.0047	0.0013	0.0003
Unemploy	29,833	-0.0109	0.0031	0.0005
IMDemployment	29,833	-1.1065	0.3199	0.0005
IMDscore	29,833	-0.0046	0.0014	0.0007
LTunemp	29,834	-0.1317	0.0379	0.0005
Godchild	29,835	0.0039	0.0013	0.0024
IMDhealth	29,836	-0.0604	0.0205	0.0032
IMDcrime	29,837	-0.0618	0.0229	0.0070
Acute	29,838	-0.1464	0.0571	0.0103
IMDservices	29,838	0.0039	0.0016	0.0156
BingeDrink	29,839	-0.0081	0.0033	0.0138
TT1564	29,839	0.0092	0.0042	0.0274
Dist2PriSch	29,839	0.0726	0.0316	0.0217
TotTurbulence	29,841	0.0083	0.0046	0.0697
Densitypph	29,843	-0.0003	0.0005	0.5806
NetFlow	29,844	0.0217	0.0204	0.2892
Dist2FdShp	29,844	0.0120	0.0114	0.2898
NF1564	29,844	-0.0055	0.0190	0.7730
IMDenvironment	29,844	0.0005	0.0012	0.6726

59 There is little to commend constructing a new model and we have used precisely the same model structure to explain variations in the likelihood that respondents to the *APS6* would meet the NI8 participation threshold as was used in the '1×30' model. Of course, the parameter values are different, as shown in Table 10 and Figure 3 below.

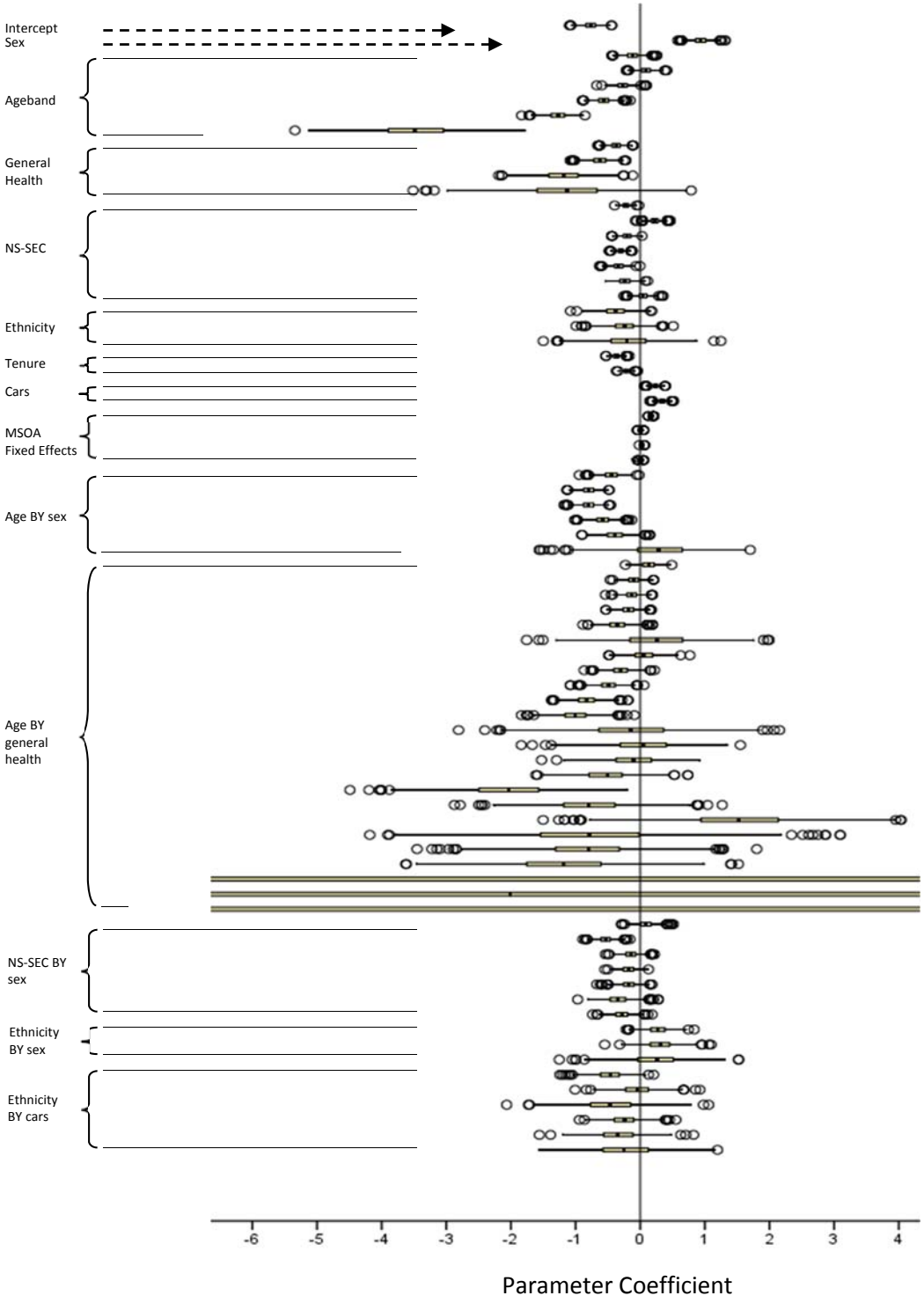
Table 10 NI8 Model fixed effect posterior means and 2.5 & 97.5% percentiles

Factor [Ref. Group]	mean	2.5%	97.5%
Constant	-0.763	-0.980	-0.548
Ageband [16-24]			
25-34	-0.122	-0.347	0.104
35-49	0.088	-0.112	0.288
50-64	-0.268	-0.473	-0.063
65-74	-0.563	-0.794	-0.333
75-84	-1.275	-1.578	-0.967
85+	-3.462	-4.613	-2.337
Sex [Female]			
Male	0.939	0.727	1.150
General Health [Very Good]			
Good	-0.371	-0.546	-0.195
Fair	-0.623	-0.892	-0.358
Bad	-1.198	-1.854	-0.551
Very Bad	-1.151	-2.462	0.160
NS-SEC† [NS-SEC 1]			
NS-SEC 2	-0.223	-0.341	-0.106
NS-SEC 3	0.221	0.077	0.366
NS-SEC 4	-0.202	-0.368	-0.034
NS-SEC 5	-0.304	-0.424	-0.185
NS-SEC 6	-0.334	-0.515	-0.153
NS-SEC 7	-0.242	-0.470	-0.013
NS-SEC 8	0.038	-0.138	0.213
Ethnicity [White or mixed]			
Black	-0.385	-0.785	0.015
Asian	-0.249	-0.678	0.183
Other	-0.188	-0.925	0.551
Tenure [Owned]			
Social Rented	-0.369	-0.480	-0.258
Other rented & other	-0.212	-0.306	-0.119
Cars [No cars]			
1 car	0.234	0.133	0.334
2+ cars	0.336	0.228	0.445
MSOA Fixed Effects			
PCA Factor 1	0.116	0.084	0.147
PCA Factor 1	-0.025	-0.061	0.010
PCA Factor 1	0.033	0.003	0.063
PCA Factor 1	0.001	-0.031	0.032
Ageband:sex			
25-34:male	-0.443	-0.689	-0.198
35-49:male	-0.796	-1.010	-0.577
50-64:male	-0.798	-1.020	-0.576
65-74:male	-0.582	-0.836	-0.331
75-84:male	-0.394	-0.735	-0.054
85+:male	0.304	-0.713	1.320

Factor [Ref. Group]	mean	2.5%	97.5%
Ageband:general health			
25-34:good health	0.127	-0.115	0.371
35-49:good health	-0.093	-0.299	0.115
50-64:good health	-0.125	-0.338	0.088
65-74:good health	-0.176	-0.420	0.068
75-84:good health	-0.357	-0.710	-0.006
85+:good health	0.238	-0.916	1.393
25-34:fair health	0.057	-0.311	0.425
35-49:fair health	-0.305	-0.621	0.009
50-64:fair health	-0.490	-0.811	-0.172
65-74:fair health	-0.837	-1.206	-0.468
75-84:fair health	-1.018	-1.517	-0.519
85+:fair health	-0.139	-1.617	1.336
25-34:bad health	0.026	-0.940	0.990
35-49:bad health	-0.122	-0.868	0.631
50-64:bad health	-0.517	-1.285	0.255
65-74:bad health	-2.041	-3.356	-0.724
75-84:bad health	-0.778	-1.981	0.429
85+:bad health	1.515	-0.288	3.333
25-34:v. bad health	-0.828	-3.244	1.590
35-49:v. bad health	-0.801	-2.355	0.756
50-64:v. bad health	-1.146	-2.746	0.445
65-74:v. bad health	-12.760	-40.054	58.321
75-84:v. bad health	-12.130	-55.064	58.894
85+:v. bad health	-10.680	-79.245	74.224
NS-SEC:sex			
NS-SEC 2:male	0.094	-0.138	0.327
NS-SEC 3:male	-0.538	-0.737	-0.336
NS-SEC 4:male	-0.144	-0.357	0.071
NS-SEC 5:male	-0.182	-0.397	0.030
NS-SEC 6:male	-0.184	-0.437	0.069
NS-SEC 7:male	-0.336	-0.700	0.026
NS-SEC 8:male	-0.275	-0.532	-0.021
Ethnicity:sex Ethnicity:sex			
Black:male	0.272	-0.051	0.593
Asian:male	0.299	-0.135	0.736
Other:male	0.245	-0.526	1.013
Ethnicity:Access to cars			
Black:1 car	-0.452	-0.885	-0.023
Asian:1 car	-0.064	-0.555	0.428
Other:1 car	-0.436	-1.332	0.457
Black:2+cars	-0.246	-0.686	0.194
Asian:2+cars	-0.342	-0.963	0.276
Other:2+cars	-0.213	-1.200	0.777

† Where NS-SEC 1 = 'Managerial & Professional'; NS-SEC 2 = 'Intermediate'; NS-SEC 3 = 'Small employers & own account workers'; NS-SEC 4 = 'Lower supervisory and technical'; NS-SEC 5 = 'Semi-routine'; NS-SEC 6 = 'Routines'; NS-SEC 7 = 'Never worked or LT Unemployed'; and NS-SEC 8 = 'Not classified (FT student, not stated or poorly described)'

Figure 3 NI8 Model fixed effect parameter estimates



60 Having derived posterior distributions for each model parameter, the next stage is to apply them to individuals and to aggregate and summarize the resulting responses to MSOA-level. The method ('microsimulation') by which we obtain a population of individual adults in each English MSOA is described in the next section, but the key here is that by applying the appropriate model parameter posterior distributions to all individuals in all MSOAs (each of whom has their particular characteristics, including in which MSOA they reside) we generate corresponding posterior distributions describing the likelihood that those person types in those MSOAs will hit the '1×30' or NI8 participation thresholds. It is on the basis of these posteriors (i.e. sets of independent estimates) that we then derive summary point estimates and 95% CIs of the number of people participating in sports or other forms of active recreation in each MSOA. Whilst theoretically straightforward, this is, as one might imagine, computationally demanding with databases of 60Gb and more. The purpose, however, is to retain in the final local area estimates all uncertainty that existed in the originating model.

61 It is important here to emphasise the essential nature of small area estimation and how this must affect any interpretation of the estimates produced by applying the model parameters to local covariate data. If a particular group of people, nationally, is found to have particularly high participation rates, then it is assumed, unless there is evidence to the contrary, that this will apply to all local areas. The multilevel nature of the model will capture whether this relationship is mediated by any MSOA-level covariate data, but the same principle applies – local area estimates are produced on the basis of modelled relationships derived from an analysis of the dataset as a whole.

62 It is thus possible that unknown (and perhaps unknowable) characteristics of particular places will result in anomalous levels of participation that will not be picked up by small area estimation – anomalous in the sense that they are out-of-line with what might be expected given what is known more widely about the relationship between, on the one hand, individual- and MSOA-level characteristics and, on the other, rates of participation in sports and other forms of active recreation. Such cannot be captured by small area estimation unless a reasonable sample of individuals is surveyed from all MSOAs in the country. The *APS6*, although large, provides an average of only 21 persons per MSOA for 6,767 of the 6,791 MSOAs for which estimates are being produced.

63 The results from small area estimation should not be used as the basis of a 'performance league table'. The approach is better suited as a mechanism for identifying and highlighting those areas where participation rates are likely to be relatively poor and where, therefore, it is likely that policy initiatives would be best directed.

4 APS6 Small Area Estimation: Microsimulation and Prediction

64 In essence, small area estimation rests on the use of two distinct sets of data. On the one hand are the survey data employed, as detailed above, to derive multilevel models which describe how a dependent variable responds to a series of individual- and area-level predictor variables. As these data have been collected about adults living in households this defines the population to which we can apply our models. We cannot, for example, say anything about children or, importantly, about people living in communal establishments. Any limitations or other issues concerning *APS6* data will have been considered elsewhere, though it is worth noting that Small Area Estimation is not as sensitive to sampling issues as traditional survey-based estimation methodologies. Obtaining as diverse a sample population as possible – given sample size – is the key criterion and, in this respect, the *APS6* provides an entirely adequate basis for modelling.

65 On the other hand are those data used to define the socio-demographic composition of local areas, along with area-level variables that equate to those used in the upper-level of the model. The survey and local covariate data must correspond as the goal is to apply posterior distributions for the various factors used in the model to *individuals* in local areas. Thus, to take a hypothetical (and unrealistically simple) example of four age and two sex categories at the individual level and a single upper-level variable, say the *2010 Index of Deprivation*, then it is necessary to establish how many people in each area are in each of the (4×2=8) age-sex categories, as well as each area's *2010 IMD* score. It is to individuals (with their age and sex characteristics) in areas (with their *2010 IMD* scores) that the modelled posterior distributions are applied in order to establish the likelihood that those individuals will meet the '1×30' or NI8 participation thresholds. By summing those likelihoods across each MSOA as a whole we derive estimates of the number of people meeting those thresholds in each MSOA.

66 This may appear relatively straightforward, but great complexity arises once 'real-world' models are constructed. The '1×30' and NI8 models both comprise five individual-level variables: sex (2 factors), ageband (7), genhealth (5), nssec (8), tenure (3), ethnic (4) and cars (3). It defines, in other words, some 2×7×5×8×3×4×3=20,160 different 'person types'. The challenge is that it is necessary to determine how many of each person type there are in each of the 6,791 MSOAs for which predictions are required.

67 Unfortunately, information concerning the detailed composition of MSOA populations simply does not exist. It is, however, possible to use what is known about the aggregate characteristics of any given population (i.e. how many males and females, how many people in each ageband, how many in each NS-SEC category, etc.) in order to deduce – or *microsimulate* – the likely number of people with each unique combination of characteristics (for instance the number of 16-24 year-old males in managerial or professional occupations; the number of 16-24 year-old females in managerial or professional occupations; and so on). The defining characteristic of a successfully microsimulated population is that, when

aggregated, it will match in all respects what is known about the overall characteristics of that population.

68 A number of 2011 Census tables have been published which provide information on partial joint distributions (i.e. two-, three and four-way tables) which can be combined to estimate the composition of the otherwise unknown 'full joint distribution' we require for each MSOA. This 'full joint distribution' describes the number of individuals in a population with each unique combination of characteristics, which in the present case involves the 20,160 person types described by the '1×30' and NI8 models. These models were defined partly to minimise the number of tables used, which are listed in Table 11 below.

Table 11 Census Tables used to microsimulate MSOA Populations

2011 Census Detailed Characteristics (DC) Tables[17]	
DC3601EW	General health by NS-SeC by sex by age (Usual residents aged 16 +)
DC6303EWR	NS-SeC by general health by sex by age (regional) (Usual residents aged 16 +)
DC6206EW	NS-SeC by ethnic group by sex by age (Usual residents aged 16 +)
DC1104EW	Residence type by sex by age (All usual residents)
DC4203EW	Tenure by car/van availability by ethnic group (All usual residents in households)
DC3302EW	Health status by sex by age (All usual residents in households)

69 A technique known as 'Iterative Proportional Fitting' (IPF) has long been used as a method of combining marginal distributions (and two- and three-way joint distributions) to derive the required full joint distribution[18]. We have adapted this approach to incorporate the *pro rata* splitting of agebands and other categories where appropriate. Until 2011 census microdata is published (updating the 2001 census 3% Sample of Anonymised Records (SARs) data, which is available at regional level, and/or the less comprehensive 5% Small Area Microdata (SAM), which is available down to local authority level) there are no grounds upon which to constrain the microsimulation. A detailed four-way regional table (DC6303EWR) is used, but this provides only a partial substitute for the detailed information on the composition of regional and local authority populations that can be extracted from microdata. As noted in Section 3 above, this means there is a higher (but unquantifiable) degree of uncertainty in the microsimulation than would otherwise be the case. This particularly applies estimates of patterns of tenure and car ownership relative to age, sex, general health and occupational social class (NS-

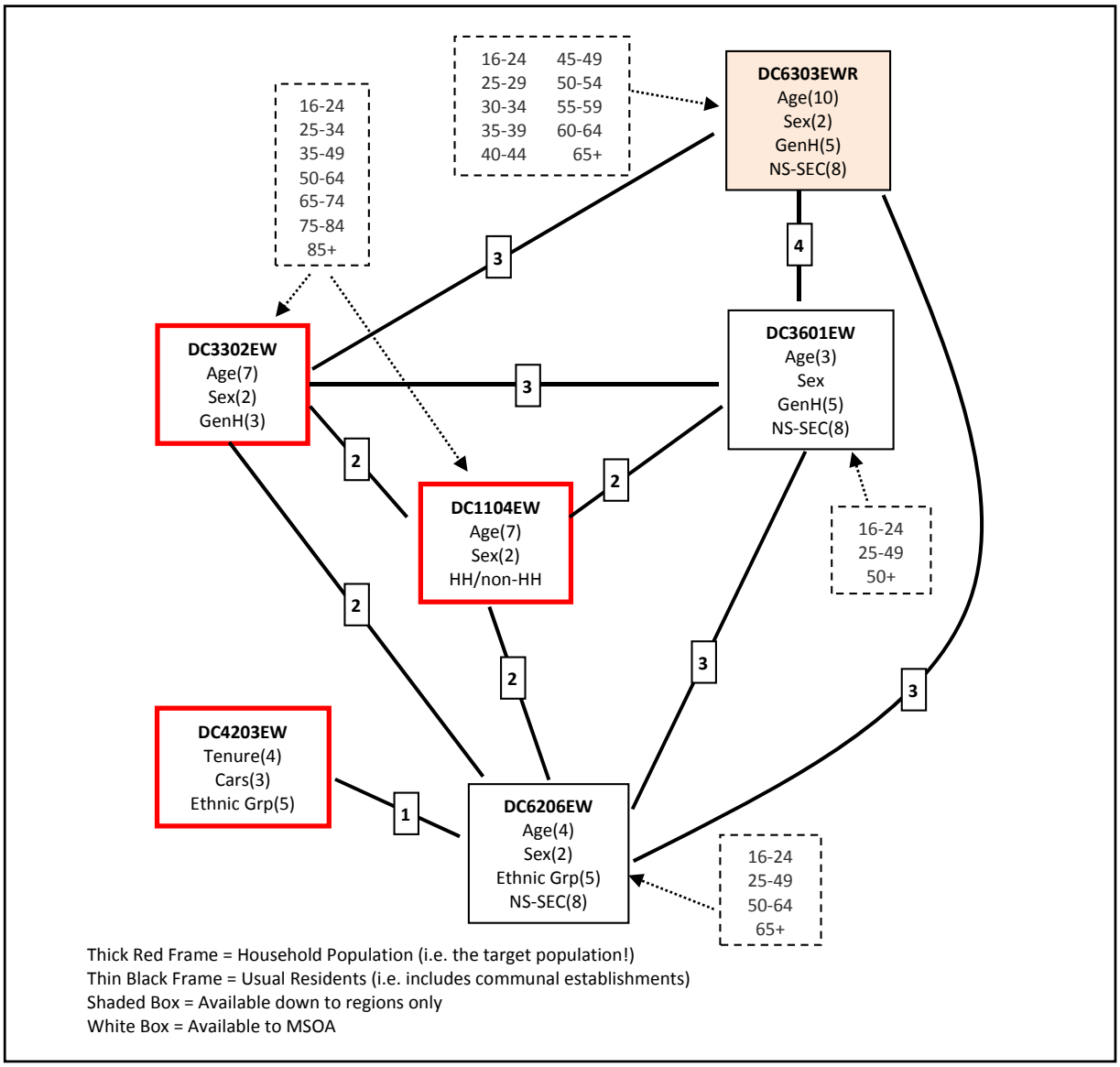
[17] *2011 Census Data for England and Wales on Nomis*. (Online data available at http://www.nomisweb.co.uk/census/2011) [Accessed 30/10/2013.]

[18] Deming, W.E. and Stephan, F.F., 1940, "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known", *Annals of Mathematical Statistics*, Vol. 11, pp427-444; Fienberg, S.E., 1970, "An iterative procedure for estimation in contingency tables", *Annals of Mathematical Statistics*, Vol. 41, pp907-917; Clarke, M. and Holm, E., 1987, "Microsimulation methods in human geography and planning: a review and further extensions", *Geografiska Annaler*, Vol. 69B, pp145-164; Birikin, M. and Clarke, M., 1988, "SYNTHESIS – a synthetic spatial information system for urban and regional analysis: methods and examples", *Environment and planning A*, Vol. 20, pp1645-1671.

SEC). This is because, as illustrated below, the only table available describing either tenure or car ownership is poorly linked with the remaining tables.

70 Whilst not optimal, it seems to us unlikely that this will materially affect the final estimates. In any case, 2011 census microdata are not due to be published until late-2013/early-2014 and until then IPF cannot be constrained to deliver an optimal microsimulation.

Figure 4 Map of Census Tables used to microsimulate MSOA populations



5 Summary of Results

71 This study is not concerned with an analysis of the small area estimates themselves, nor with what they may tell us about how or why participation in sport and active recreation varies at the very local level, but it is important to determine whether or not the estimates exhibit 'face validity'. The degree to which the results are both internally consistent and align with what might be expected.

72 In large measure, this assessment of face validity will have to lie with those expert in the field and on the basis of an examination of the detailed data made available in the accompanying spreadsheets, but we can here attempt to summarise the findings and draw attention to a number of key features. To that end this section includes a series of maps plotting our estimates of the proportion of people meeting the '1×30' and NI8 participation targets. The maps have all been designed to emphasise areas of high participation rates, i.e. areas are most strongly shaded where participation rates are highest.

73 The MSOA maps exhibit impressive granularity, though it worth repeating the observations made above about the nature of these *modelled* estimates. Small Area Estimation 'pools' evidence on the basis of modelled relationships that have been derived from an analysis of the dataset as a whole. It is best suited, therefore, as a mechanism for identifying those areas where participation rates are *likely* to be relatively poor and where, therefore, it is *likely* that policy initiatives would be best directed. It cannot capture genuinely anomalous local variations from the norm.

74 Yet the analysis does, as illustrated by Figure 5 to Figure 12, serve to highlight some important patterns with respect to participation rates at the local level. First, as might be expected, there are huge variations at the very local level. For instance, although the overall proportion of adults meeting the '1×30' participation rate varies from 33.6%(Tendring) to 53.8% (Wandsworth) at local authority level, at MSOA-level the proportion ranges from 24.9% (Tendring 018) to 69.3% (Newcastle upon Tyne 013). This pattern is repeated with respect to NI8 participation rates, and it is clear that areas of low participation can be very localised. Small Area Estimation provides an invaluable means of highlighting and quantifying this local dimension.

75 Second, and continuing this theme, it is clear that there are some very marked variations in participation rates within individual local authority areas. Unsurprisingly, such variation tends to be most marked in large urban authorities, such as Newcastle upon Tyne which includes within it individual MSOAs with rates varying from 25.9% to 69.3%. More rural areas are not exempt from this, with many rural local authorities showing similar diversity; such as Northumberland which contains individual MSOAs with rates which vary from 28.2% to 54.9%. The perspective afforded by Small Area Estimation strongly suggests that any policies aimed at addressing low levels of sporting participation may need to be spatially fine grained.

76 Having observed such variations, the third point is to note that it can, in fact, be very misleading to focus on overall rates of participation. For instance, Newcastle upon Type 013 (the MSOA noted above with the highest proportion of adults meeting the '1×30' threshold) covers an area of student accommodation around the university and no less than 64.9% of adults are aged 16-24. Parameter estimates for both the '1×30' and NI8 models (Table 7 and Table 10 above respectively) emphasise the huge importance of age in determining rates of participation. This is hardly surprising, of course, but it follows that local rates of participation will respond strongly to variations in the demographic structure of local areas. These are more pronounced than is often recognised. Thus at Local Authority level the proportion of people aged, for instance, 75+ varies from 2.9% in Tower Hamlets (London) to 16.0% in Christchurch (Dorset). At MSOA level the variation is even more marked, with the proportion of people aged 75+ or more varying from less than 1% in a number of MSOAs (presumably largely associated with student populations) in Leeds, Liverpool, Manchester, Leicester, Birmingham, Swindon, Thurrock and Tower Hamlets to as much as 28.8% in East Devon 012.

77 Fourth, the detailed MSOA maps (along, indeed, with the LA-level maps) draw attention to the fact that low rates of participation are not a particularly urban phenomenon. Thus even when one focuses on a particular age-sex cohort it is around Lincolnshire, the Wash and northern East Anglia, as well as across many northern and western counties, that participation rates appear particularly low. In this respect it is worth looking again at Figure 1. Once again, in other words, the evidence afforded by Small Area Estimation adds significantly to our understanding of how participation rates vary nationally as well as locally.

78 Our final comment should perhaps be directed at the issue of estimate uncertainty. Developing a methodology that could capture estimate uncertainty was a key design consideration for this project, but it should be noted that what we have termed 'model uncertainty' – i.e. the quantifiable level of confidence we have in our modelled estimates *given the data* – is likely to underestimate the overall level of *model+data* uncertainty. We include in the output files (described in Section 6 below) the 95% CIs for each and every estimate of the number and proportion of people meeting the '1×30' and NI8 participation thresholds, but these are estimates of *model* uncertainty only and must thus be interpreted accordingly. For instance, an additional (small but essentially unquantifiable) degree of uncertainty surrounds these CIs simply because of the uncertainty surrounding the microsimulation of the socio-economic composition of local populations – which has been produced using IPF and *pro rata* microsimulation unconstrained by 2011 microdata. Whilst we have confidence in the broad thrust of the analysis, it is possible that in some places unique circumstances give rise to levels of participation that are higher or lower than predicted.

Figure 5 MSOA-level '1×30' participation rates; Adults

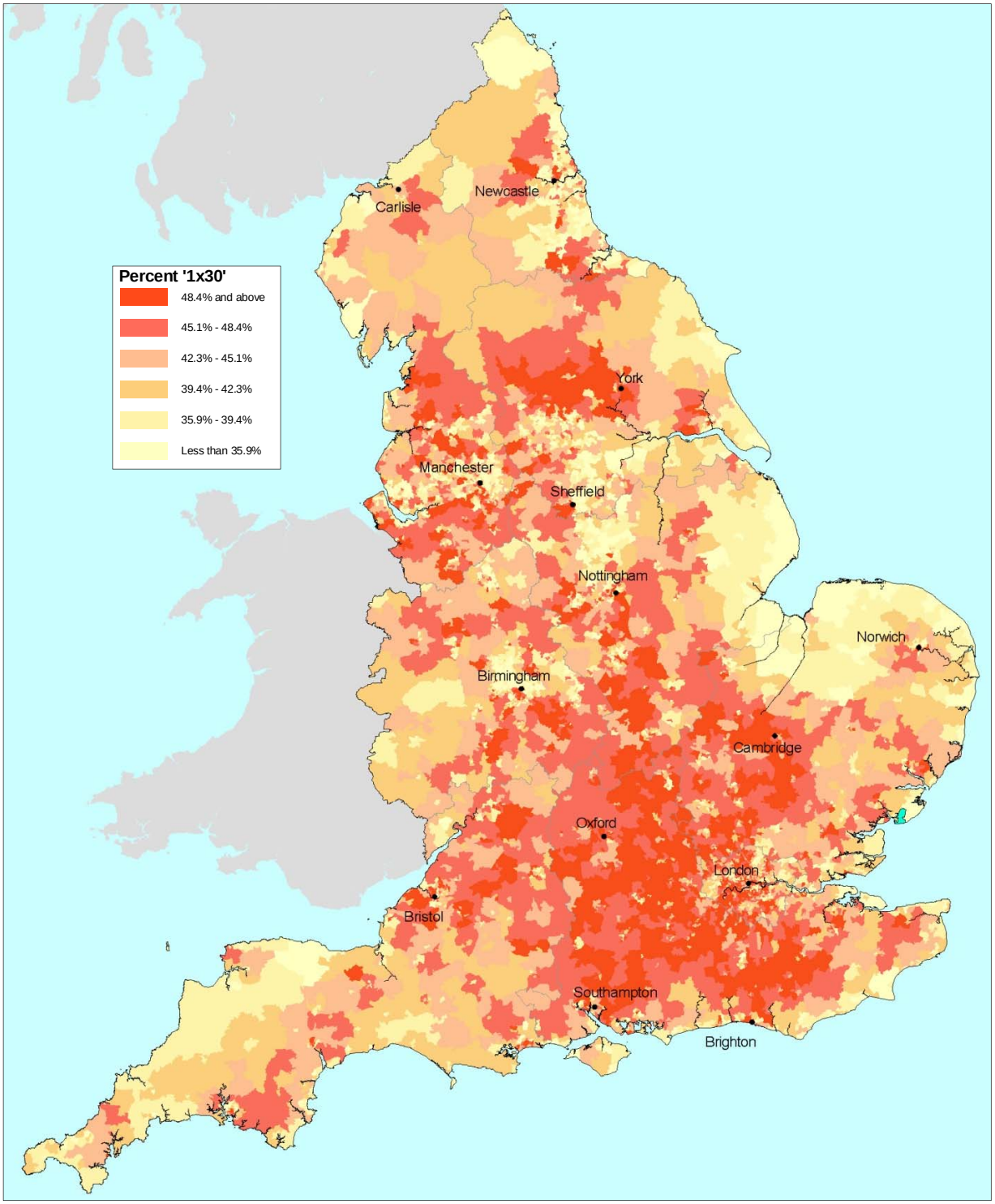


Figure 6 MSOA-level NI8 Participation Rates: All Adults

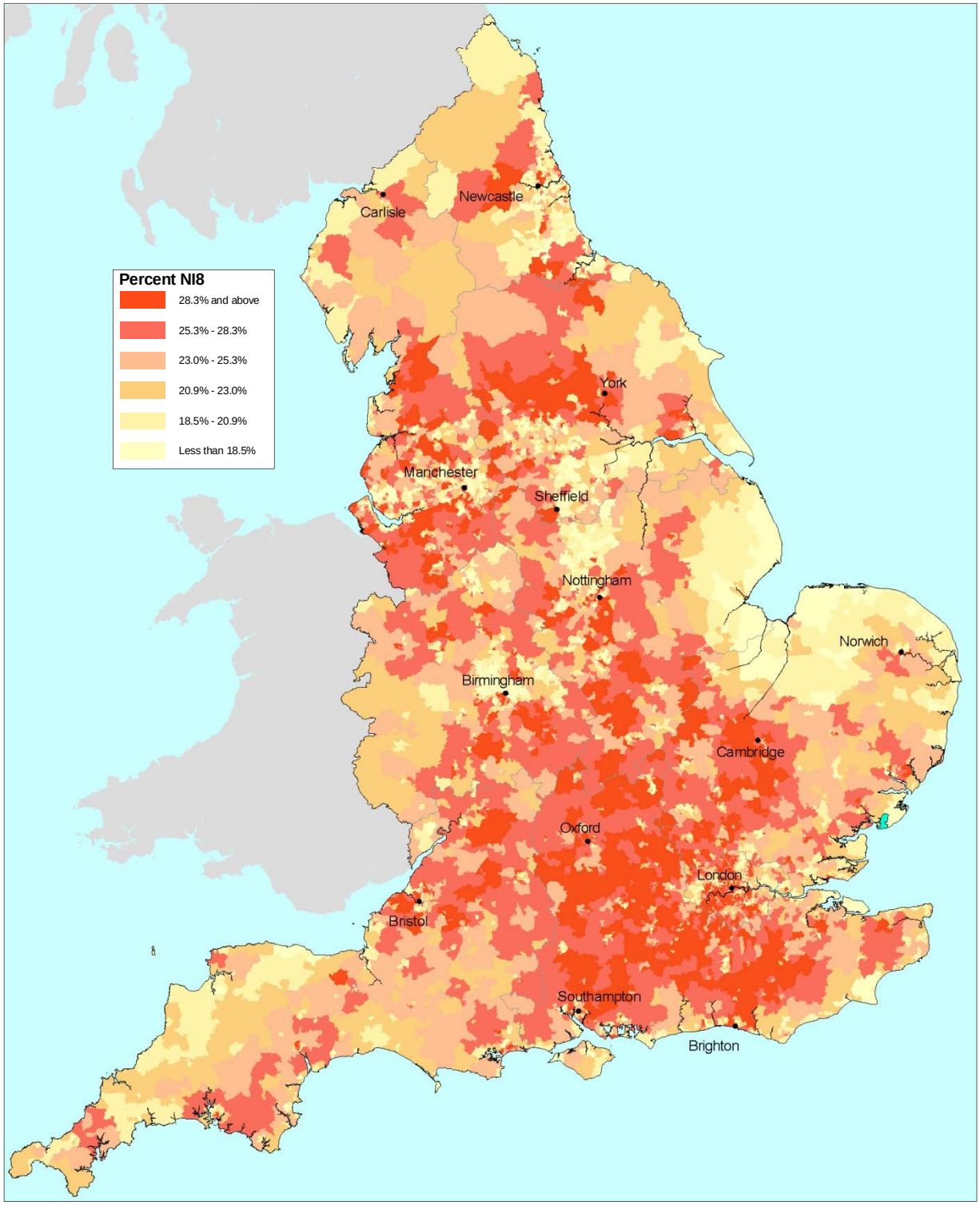


Figure 7 Local Authority Level '1×30' Participation Rates; All Adults - REMOVE

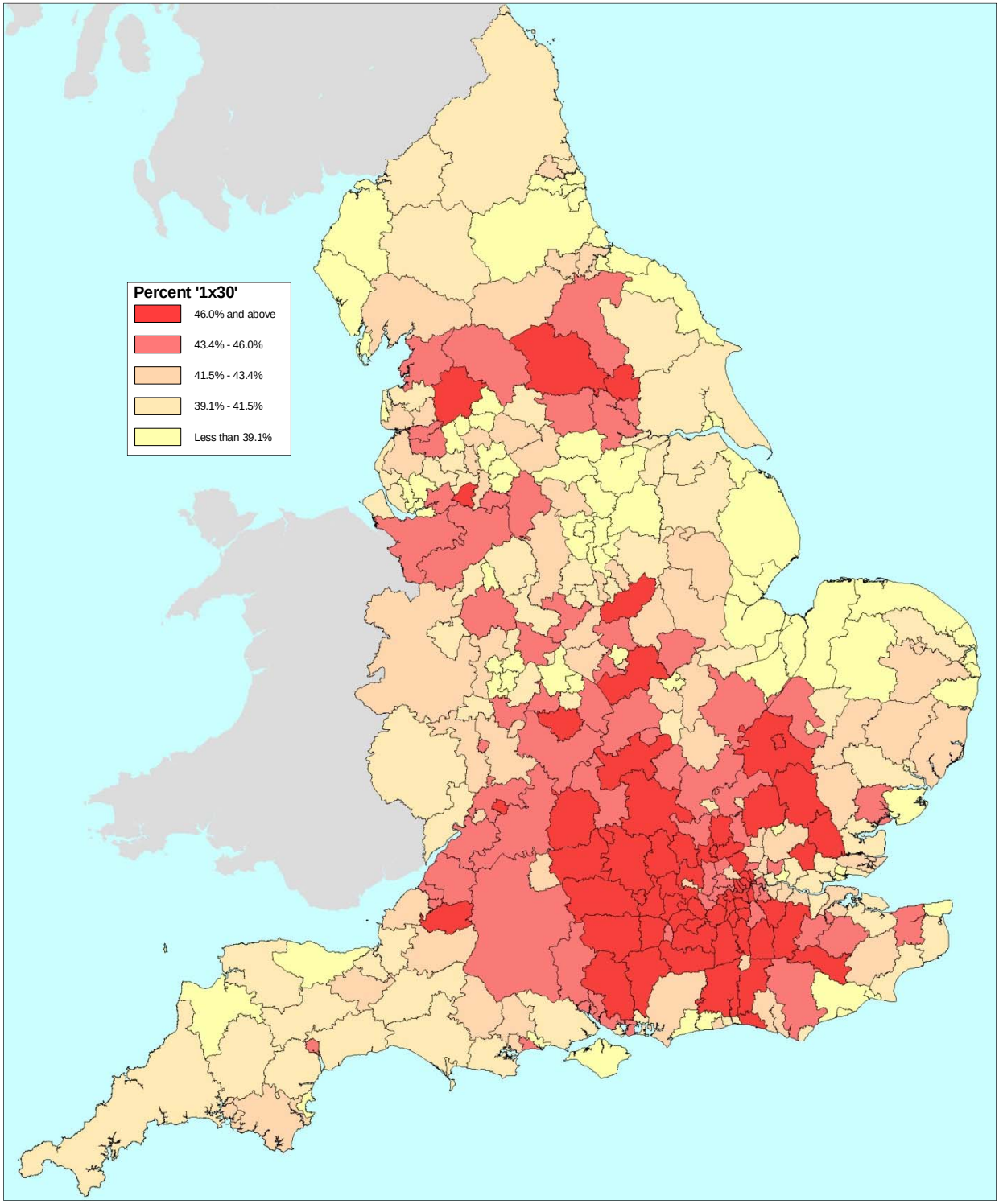


Figure 8 Local Authority Level NI8 Participation Rates; All Adults - REMOVE

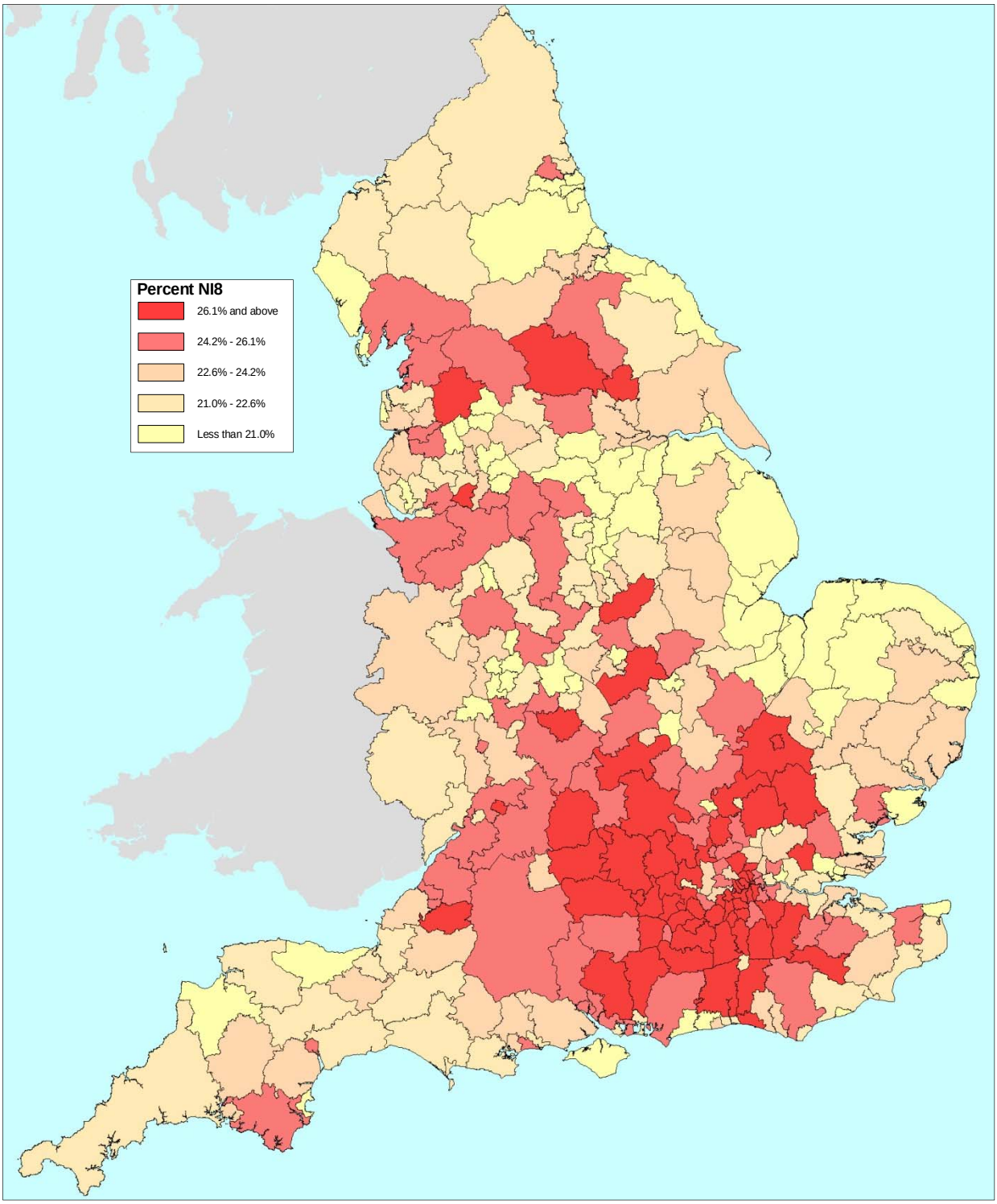


Figure 9 MSOA-level '1×30' participation rates; Males 25-34

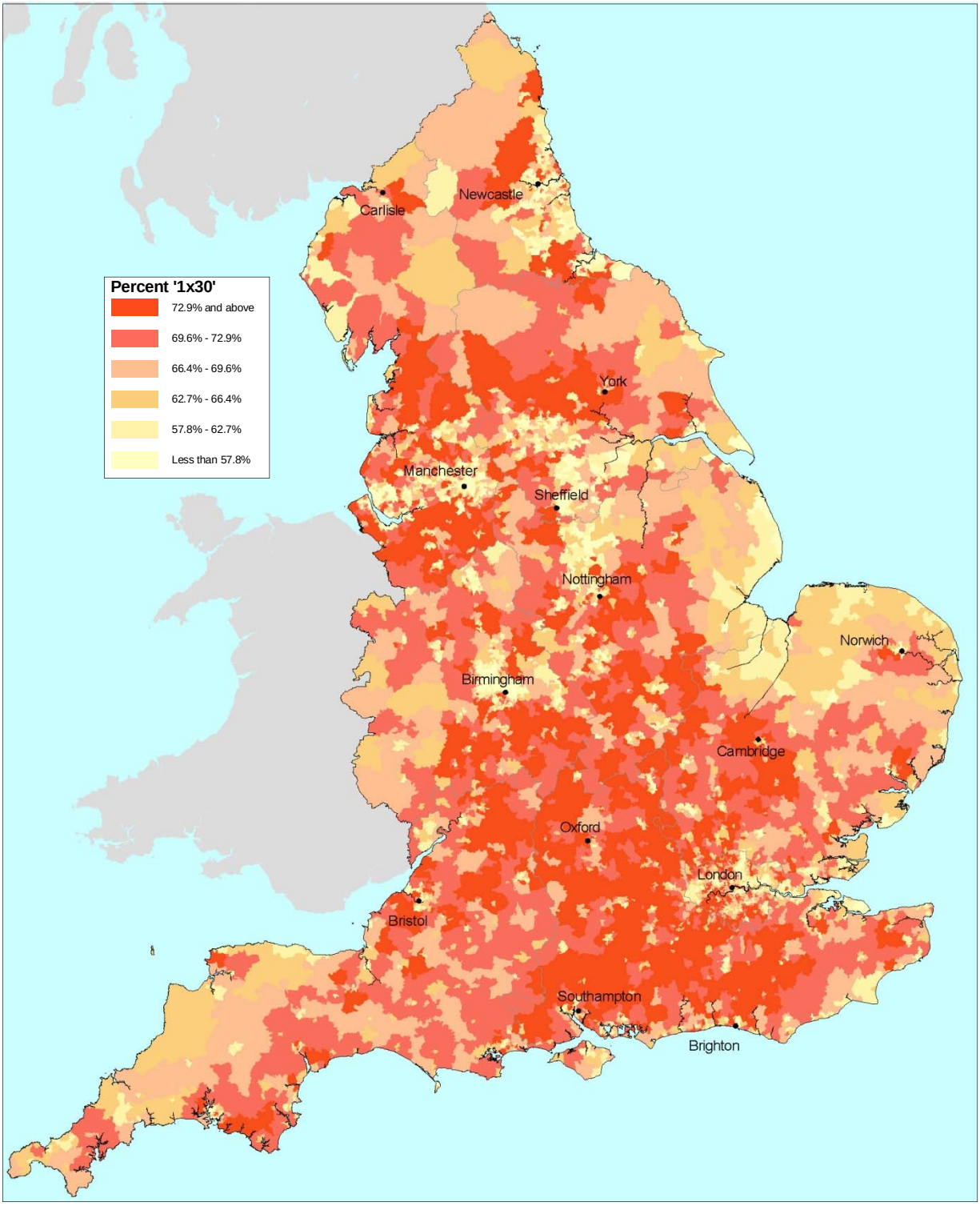


Figure 10 MSOA-level NI8 Participation Rates: Males 25-34

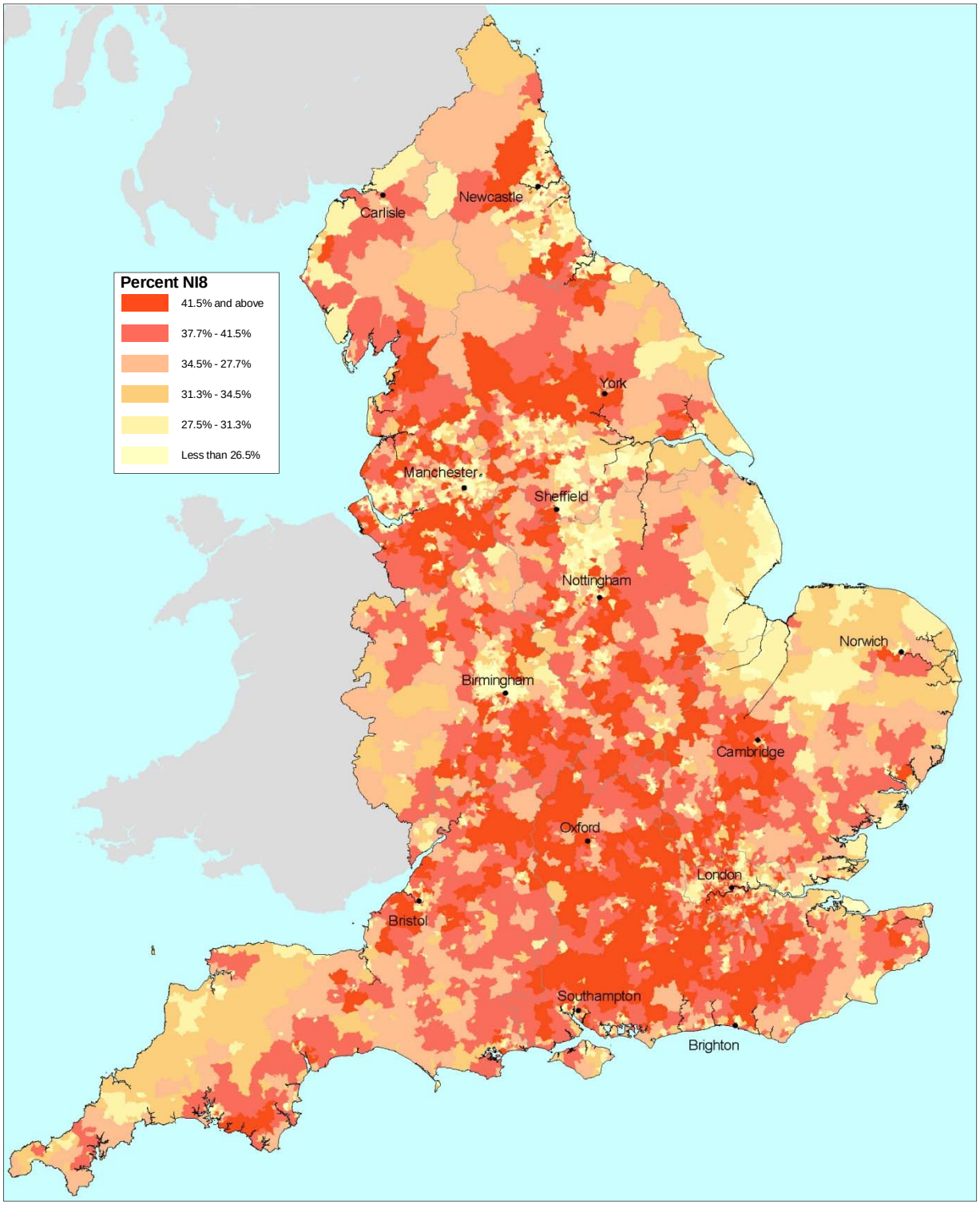


Figure 11 Local Authority Level '1×30' Participation Rates; Males 25-34 REMOVE

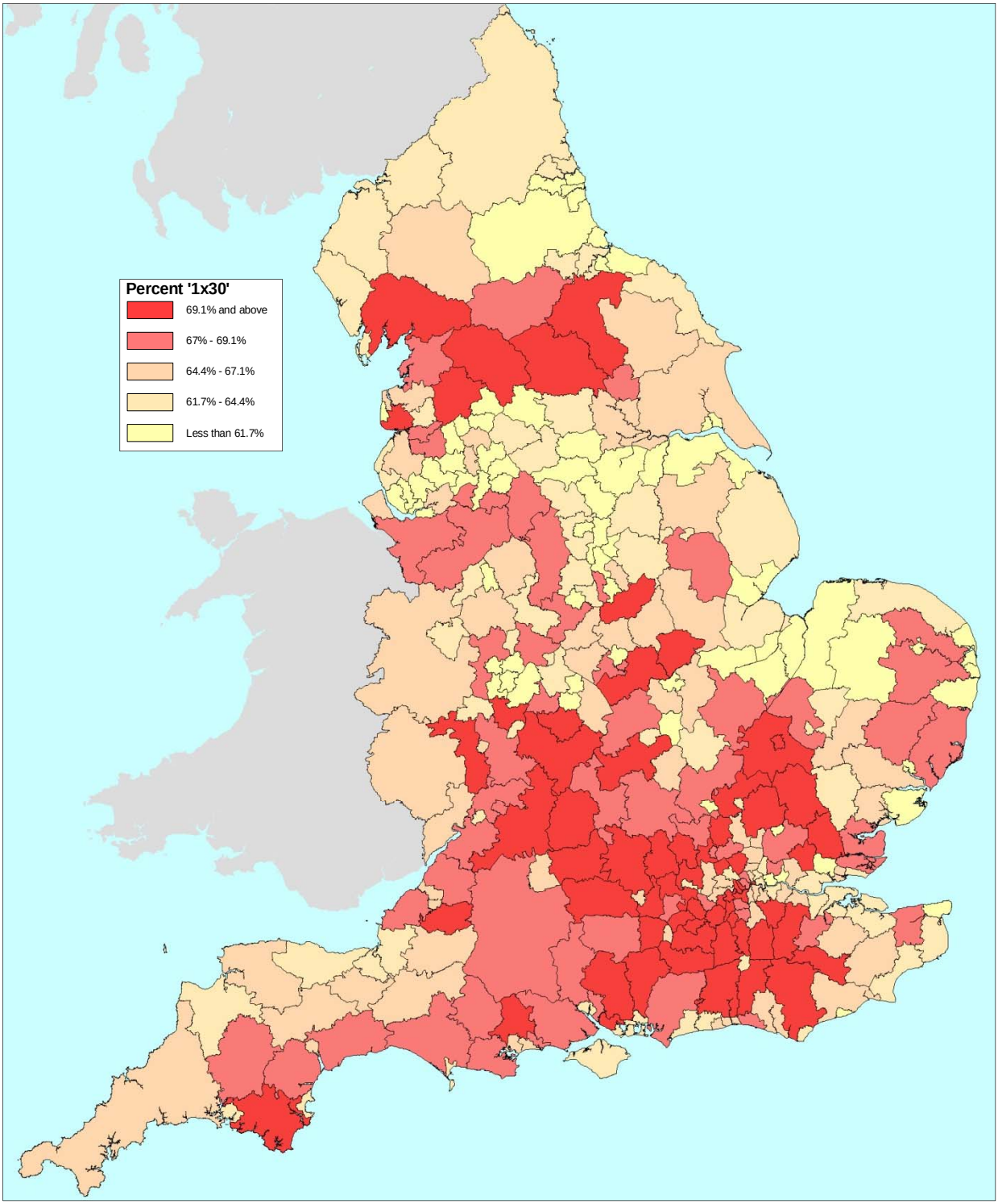
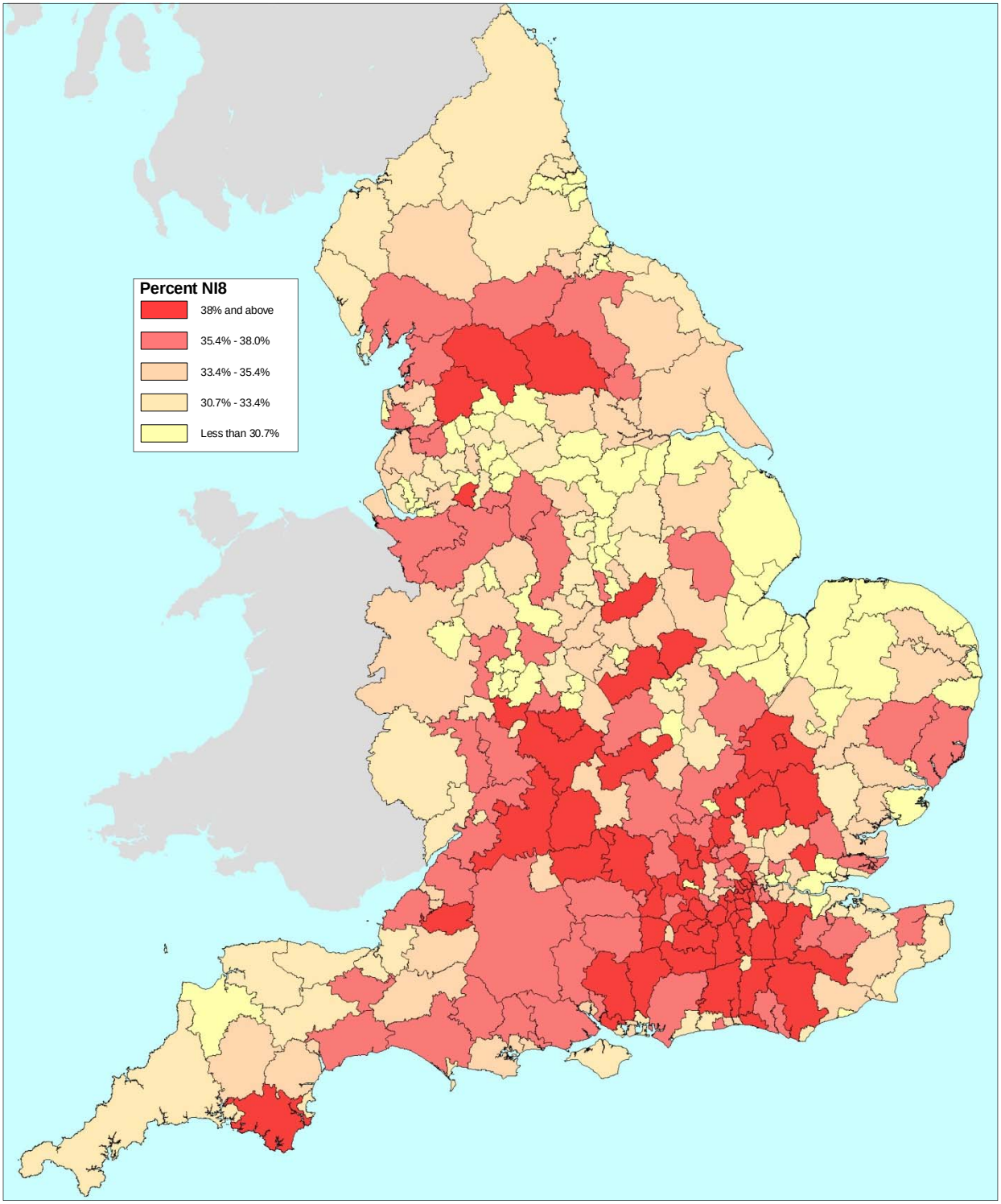


Figure 12 Local Authority Level NI8 Participation Rates; Males 25-34 REMOVE



6 Summary Guide to Local Area Prediction Spreadsheet Files

79 The data are supplied in four Excel workbooks, each of which contains three spreadsheets. The four workbooks are as follows:

(a) MSOA SAE Estimates 1×30.xlsx

(b) MSOA SAE Estimates NI8.xlsx

(c) LA SAE Estimates 1×30.xlsx

(d) LA SAE Estimates NI8.xlsx

80 The three spreadsheets in each workbook contain tables which present, respectively, (a) estimated counts, (b) population denominators and (c) participation rates for the following fourteen age-sex cohorts, plus for the adult population as a whole. For each cohort we provide a point (mean) estimate for the posterior distribution, an upper 95% credible interval estimate for the distribution, and a lower 95% credible interval estimate for the distribution. Each table thus contains the following columns:

Columns	Cohort	Data
1	MSOA11CD/LAD11NM	MSOA Standard Code
2	MSOA11NM/LAD11NM	MSOA Standard Name
3-5	Females, 16-24	F_16-24m; F_16-24u; F_16-24l
6-8	Females, 25-34	F_25-34m; F_25-34u; F_25-34l
9-11	Females, 35-49	F_35-49m; F_35-49u; F_35-49l
12-14	Females, 50-64	F_50-64m; F_50-64u; F_50-64l
15-17	Females, 65-74	F_65-74m; F_65-74u; F_65-74l
18-20	Females, 75-84	F_75-84m; F_75-84u; F_75-84l
21-23	Females, 85 and over	F_85+m; F_85+u; F_85+l
24-26	Males, 16-24	M_16-24m; M_16-24u; M_16-24l
27-29	Males, 25-34	M_25-34m; M_25-34u; M_25-34l
30-32	Males, 35-49	M_35-49m; M_35-49u; M_35-49l
33-35	Males, 50-64	M_50-64m; M_50-64u; M_50-64l
36-38	Males, 65-74	M_65-74m; M_65-74u; M_65-74l
39-41	Males, 75-84	M_75-84m; M_75-84u; M_75-84l
42-44	Males, 85 and over	M_85+m; M_85+u; M_85+l
45-47	Total/Overall	Totalm; Totalu; Totall

7 Appendix Supplementary Data Sources

The '1×30' and NI8 models use individual- and MSOA-level data to predict whether or not individuals in the *APS6* dataset (n=163,420) claim to have met the respective participation thresholds. All individual-level data are drawn from the *APS6* itself, but MSOA-level variables must be drawn from supplementary sources describing various characteristics of the MSOAs in which individuals reside. We have information on the MSOAs in which 136,996 (83.8%) of respondents live.

This section briefly describes the MSOA-level data made available (though not necessarily used) in the '1×30' and NI8 models. As described in Section 3, in fact a subset of the data was presented to Principal Component Analysis, and four factors (explaining 75% of the variability of the variables) were ultimately included in the model. The variables included in the PCA were:

Table 12 Variables included in Principal Components Analysis

Name	Variable Description
Acute	Age standardied rate of emergency admissions to hospital
BingeDrink	Modelled estimate of % binge drinkers
Densitypph	Density (persons per hectare)
Dist2PriSch	Average distance (km) to Primary School (OA centroid to school)
GCSE	% students achieving 5+ GCSEs (incl Maths & English)
goodchild	% 5 year olds with good development
hhpov	% pop in households receiving means-tested benefits
L3plusQuals	% adults with L3 qualifications or higher (2011 Census QS501EW)
L4plusQuals	% adults with L4 qualifications or higher (2011 Census QS501EW)
LTunemp	% working age pop unemployed for 12 months or more
NetTurbulence	Net GP Population Turbulence (in-out/pop)
Not_HE	% people under 21 not entering Higher Education
NT1564	Age15-64 Net GP Population Turbulence (in-out/pop)
Obese	Modelled estimate of % obese
Smoking	Modelled estimate of % smokers
TotTurbulence	Total GP Population Turbulence (in+out/pop)
TT1564	Age15-64 Total GP Population Turbulence (in+out/pop)
Veg5aday	Modelled estimate of % achieving 5-a-day veg and fruit

These variables, along with all other MSOA-level variables considered during the specification of the '1×30' and NI8 models are detailed below. In addition, there is a final brief section describing the local authority level variables which were considered, although eventually rejected.

MSOA-level Sources and Data

Super Output Areas were designed to improve the reporting of small area statistics and are built up from groups of Output Areas. Statistics for Lower Layer Super Output Areas (LSOA) and Middle Layer Super Output Areas (MSOA) were originally released in 2004 for England and Wales. Maintaining continuity was a guiding principle behind the establishment of LSOAs and MSOAs, but in some areas significant population changes between 2001 and 2011 meant that minimum and maximum thresholds for these units were breached. Simplistically, where populations have become too big, the LSOA/MSOA has been split into two or more areas; where populations have become too small the LSOA/MSOA has been merged with an adjacent one. More problematically, in a few areas a more complex reorganisation has been implemented. The vast majority of MSOAs (6,640 of the original 6,781) are, however, unaffected by these changes.

The reorganisation of LSOAs and MSOAs for reporting 2011 Census data has resulted in 6,791 English MSOAs. Almost all (99.6%) are represented in the *APS6* dataset (n=6,767 with, on average, 24.1 persons per represented MSOA). The problem is that, because the new LSOA/MSOA geography is so recent, most data were collected and processed using pre-2011 LSOAs and MSOAs. The widely used Index of Multiple Deprivation (2010) is a case in point.

In some cases the data have been reattributed to the new MSOA geography (such as the data published by Public Health England), but where the data are only available for pre-2011 MSOAs we have used population-weighted attribution techniques to attach the data to the new geography. This replicates the method used by Public Health England and is based on linking output areas (OAs) to both pre- and post-2011 MSOA, and weighting the link between the two MSOA geographies relative to the 2011 OA adult (16+) populations. We note in the following paragraphs whether or not the data have been processed in this way.

2011 Census data (NOMIS):

https://www.gov.uk/government/publications/2011-rural-urban-classification

MSOA_Densitypph	Density (persons per hectare) (QS102EW)
MSOA_L4plusQuals	% adults with Level 4 qualifications or higher (QS501EW)
MSOA_L3plusQuals	% adults with Level 3 qualifications or higher (QS501EW)

Notes: Drawn from the 2011 Census, these variables are definitionally straightforward and, having been derived from 2011 Census data, relate to 2001 MSOAs. The latter two variables were chosen specifically in the hope of capturing a potentially important social dimension that is notably absent from the individual-level data, namely educational status. It proved impossible to relate the *APS6* classification of educational qualifications with that used in the 2011 Census. These variables may thus to some extent act as an area-level proxy for this individual-level effect, as well as provide more genuinely areal information about local educational environment.

Public Health England Local Health Indicators

http://www.localhealth.org.uk

MSOA_goodchild	% 5 year olds with good development
MSOA_childpov	% children in households in poverty
MSOA_hhpov	% pop in households receiving means-tested benefits
MSOA_GCSE	% students achieving 5+ GCSEs (incl Maths & English)
MSOA_LTunemp	% working age pop unemployed for 12 months or more

Notes: These data were produced for 2011 MSOAs level by Public Health England as Small Area Indicators for Joint Strategic Needs Assessment. Full metadata on the variables is available[19]. All five indicators were originally produced for pre-2011 MSOAs. As described in the metadata, Public Health England used a population-weighted methodology to assign the original figures to 2011 MSOAs.

Whilst these variables will all stand as proxies for a wider set of local social, cultural and economic conditions, their specific definitions are as follows:

MSOA_goodchild The percent of foundation stage pupils in the 2010-11 academic year with a good level of development: i.e. achieving at least 78 points across all 13 Early Years Foundation Stage Profile scales (including a minimum number in particular areas of learning and development).

MSOA_childpov The percent of children in income-deprived households (mid-2009); where such households are defined as either a) receiving IS/JSA-IB/PC or b) not in receipt of these benefits but in receipt of WTC/CTC with an equivalised income below 60 per cent of the national median before housing costs.

MSOA_hhpov The percent population (mid-2009) living in low-income families reliant on means tested benefits.

MSOA_GCSE The percent pupils at the end of Key Stage 4 in schools maintained by local authorities who achieve 5 or more GCSEs at grades A*-C (including English and Maths) or equivalent at the end of the academic year 2010/11.

MSOA_LTunemp The average monthly number of persons who, in the months April 2010-March 2011 inclusive, had been claiming jobseekers allowance for more than 12 months, divided by the population aged 16-64 in mid-2010, expressed as a percentage.

ONS MSOA Population Turnover Rates, Mid-2009 to Mid-2010

http://www.ons.gov.uk/ons/rel/ness/msoa-population-turnover-rates/mid-2009-to-mid-2010/index.html

MSOA_TotTurbulence	Total GP Population Turbulence (in+out/pop)
MSOA_NetFlow	Net GP Population Flow (in-out/pop)
MSOA_TT1564	Age15-64 Total GP Population Turbulence (in+out/pop)
MSOA_NF1564	Age15-64 Net GP Population Flow (in-out/pop)

Notes: These variables aim to catch aspects of the stability (or otherwise) of local communities. Population turbulence is defined as the total number of people leaving the patient registers between July 2009 and July 2010, plus the total number joining over the

[19] Public Health England, Local Health Metadata (no date). (Available at http://tinyurl.com/owsfsne.) [Accessed 1/11/2013.]

same period, divided by estimated mid-2110 MSOA population totals[20]. Net flow refers to the difference between in-migration and out-migration, as observed in patient registers, again expressed relative to mid-2010 population totals. The data are made available for pre-2011 MSOAs and have been attributed to 2011 MSOAs using a population weighted methodology.

ONS Healthy Lifestyle Behaviours: Model Based Estimates, 2003-05

http://www.neighbourhood.statistics.gov.uk/dissemination/Download1.do (search *'Lifestyle')*

MSOA_Smoking	Estimate of percent smokers
MSOA_BingeDrink	Estimate of percent binge drinkers
MSOA_Obese	Estimate of percent obese
MSOA_Veg5aday	Estimate of percent achieving 5-a-day veg and fruit consumption

Notes: As described in the three reports which describe the production of these data[21], we are dealing here with modelled estimates based on an analysis of Health Survey for England data for 2003-5 (using a form of Small Area Estimation). The published estimates are accompanied by confidence intervals and, certainly in the case of the percent adults consuming 5 or more potions of fruit and vegetables, these can be very wide. Our interest in these data, notwithstanding that they are modelled estimates, is that they go to the heart of factors which may well be related to exercise and sporting activity. Our argument is that they are likely help distinguish between MSOAs in our predictive models. They are thus being used as 'scores' rather than explicit estimates of lifestyle behaviours around which we would need to place confidence intervals. In detail the variables are as follows:

MSOA_Smoking Adult respondents to the HSfE were defined to be current smokers if they reported that they were a "current cigarette smoker", and not a current smoker if they reported that they had "never smoked cigarettes at all", "used to smoke cigarettes occasionally" or "used to smoke cigarettes regularly".

MSOA_BingeDrink The model based estimate for the prevalence of binge drinking in adults. Adult respondents to the HSfE were defined to binge drink if men had consumed 8 or more units of alcohol or women, 6 or more units of alcohol on their heaviest drinking day in the past week.

MSOA_Obese The model based estimate for the prevalence of obesity in adults. Adult respondents to the HSfE were defined to be obese if they had a body mass index (BMI) of 30 or above.

MSOA_Veg5aday The model based estimate for the prevalence of 5+ daily Fruit & Vegetable Consumption in adults. Adult respondents to the HSfE were defined to consume Fruit & Vegetables if they had consumed 5 or more portions of fruit and vegetables on the previous day

---

[20] ONS, *Super Output Area mid-year population estimates for England and Wales , Mid-2010* (Available at http://www.ons.gov.uk/ons/rel/sape/soa-mid-year-pop-est-engl-wales-exp/mid-2010-release/index.html) [Accessed 30/1/2013.]

[21] Scholes, S., Pickering, K. & Deverill, C. (2008) *Healthy Lifestyle Behaviours: Model Based Estimates for Middle Layer Super Output Areas and Local Authorities in England, 2003-2005: Stage 1 Report / Stage 2 Report / Stage 3 Report*. The NHS Information Centre for health and social care. (Available via http://www.hscic.gov.uk/searchcatalogue?productid=PUB02479) [Accessed 30/10/2013].

English Indices of Deprivation 2010 : Scores, Domain Scores and Indicators

https://www.gov.uk/government/publications/english-indices-of-deprivation-2010

MSOA_IMDscore	Index of Multiple Deprivation (2010) Overall Score
MSOA_IMDincome	IMD2010 Income Domain
MSOA_IMDemployment	IMD2010 Employment Domain
MSOA_IMDhealth	IMD2010 Health Domain
MSOA_IMDeducation	IMD2010 Education Domain
MSOA_IMDservices	IMD2010 Barriers to Services Domain
MSOA_IMDcrime	IMD2010 Crime Domain
MSOA_IMDenvironment	IMD2010 Living Environment Domain
MSOA_Not_HE	Percent people under 21 not entering Higher Education
MSOA_Dist2FdShp	Average distance (km) to Food Shop
MSOA_Dist2PriSch	Average distance (km) to Primary School
MSOA_Acute	Age standardised hospital emergency admissions rate
MSOA_LowIncome	Percent income deprived individuals
MSOA_Unemploy	Percent employment deprived individuals

The published 2010 IMD data refers to the pre-2011 LSOAs, a small proportion of which merged, split or underwent a more complex reconfiguration to form the 32,844 English LSOAs used to describe the 2011 census. We have 'assigned' scores and the underlying indicators using the same population-weighted methodology used with respect to MSOA (i.e. by linking output areas (OAs) to both pre- and post-2011 LSOAs, and weighting the link between the two LSOA geographies relative to 2011 OA adult (16+) populations). The size of their adult populations was then used to attribute LSOA scores and underlying indicator data to MSOAs. This mapping cannot be perfect, but almost all MSOAs are entirely unaffected. More importantly, aggregating IMD2010 scores undermines the statistical methodology used to score and rank LSOAs, as detailed in the technical report accompanying the data[22].

Our argument, once again, is that the resulting 2011 MSOA scores (however imperfect) do serve to distinguish between MSOAs for the purposes of our predictive model. Using the underlying indicator data is less problematic in that the data can legitimately be aggregated from LSOAs to MSOAs. The underlying indicators are as follows::

MSOA_Not_HE The LSOA 14-17 year old population less the number of people aged under 21 in each LSOA recorded by the Higher Education Statistics Agency as entering Higher Education (2005-06 to 2008-09), divided by the LSOA 14-17 population, expressed as a percent. Four years of data were used to reduce the problems of small numbers.

MSOA_Dist2FdShp The average distance (km) from each OA centroid in the LSOA to a food shop. The definition of 'food shop' includes both larger food shops such as supermarkets as well as smaller convenience stores (approximately 16,000). Distances between OA-centriods and food shops were population weighted by 2001 OA populations.

[22] McLennan, D., *et al.* (2011) *The English Indices of Deprivation 2010*, Department for Communities and Local Government. (Available at https://www.gov.uk/government/publications/english-indices-of-deprivation-2010-technical-report.) [Accessed 12/1/2012.]

MSOA_Dist2PriSch The average distance (km) from each OA centroid in the LSOA to a primary school. All state schools classified as 'primary' were included (approximately 16,000). This includes separate infant and junior schools as well as primary schools that educate children from 5-11 years of age. Distances were population weighted as above.

MSOA_Acute An indicator of emergency admissions to hospital. The numerator is the number of hospital spells starting with admission in an emergency and lasting more than a calendar day in five year age-sex bands for 2006-07 and 2007-08. Two years of data were used to reduce the problems of small numbers. The denominator is the total population in five year age-sex bands for 2008 (each band is multiplied by three to match the three years of numerator data). Hospital admissions data were supplied by the NHS Information Centre from the Hospital Episode Statistics database, while population data were based on data supplied by the Office for National Statistics.

Local Authority Level Sources and Data

There are 326 (post-2009) local authorities in England and all respondents in the *APS6* are assigned to a local authority. With the exception of the City of London (n=73) and the Isles of Scilly (n=81) the minimum LA-level sample size was n=452. 192 of 326 LAs return a sample size of 500 or more (overall mean=501.29).

The *APS6* dataset coding for local authorities was converted to standard ONS codes (eg E08000020) and standard Local Authority names. We use these to link with the following dataset to define a series of measures relating to the density of registered sports clubs.

Sport England's Clubmark Database

http://www.clubmark.org.uk/getting-clubmark/resources-and-templates/clubmark-database

LA_Clubspp	Clubmark Clubs per 10,000 persons
LA_Clubspp1164	Clubmark Clubs per 10,000 persons aged 11-64
LA_Clubspp11plus	Clubmark Clubs per 10,000 persons aged 11 plus
LA_Clubspp1664	Clubmark Clubs per 10,000 persons aged 16-64
LA_Clubspp16plus	Clubmark Clubs per 10,000 persons aged 16 plus
LA_Clubsphec	Clubmark Clubs per 1,000 hectares

Clubs includes those 'accredited' and those 'working towards' a Clubmark accreditation. A number of different denominators were used to construct a series of potential 'supply' measures. In fact, none proved significant in individual-MSOA-LA mixed effects logistic regression models and, once we recognised how little LA-level variance was available for explanation in these models, we decided to exclude the LA-level from our models (thereby removing both LA covariate data and LA random effects from the analysis).