



---

## ESTIMATING SMALL AREA INCOME DEPRIVATION: AN ITERATIVE PROPORTIONAL FITTING APPROACH

CRESI WORKING PAPER NUMBER: 2011-02

Ben Anderson [benander@essex.ac.uk](mailto:benander@essex.ac.uk)

A version of this paper has been accepted for inclusion as a chapter in: Edwards, K., & Tanton, R. (2011) Spatial Microsimulation: A Reference Guide for Users, Springer.

**Abstract:**

Small area estimation and in particular the estimation of small area income deprivation has potential value in the development of new or alternative components of multiple deprivation indices. These new approaches enable the development of income distribution threshold based as opposed to benefit count based measures of income deprivation and so enable the alignment of regional and national measures such as the Households Below Average Income with small area measures. This paper briefly reviews a number of approaches to small area estimation before describing in some detail an iterative proportional fitting based spatial microsimulation approach. This approach is then applied to the estimation of small area HBAI rates at the small area level in Wales in 2003-5. The paper discusses the results of this approach, contrasts them with contemporary 'official' income deprivation measures for the same areas and describes a range of ways to assess the robustness of the results.

**Keywords/tags:**

spatial microsimulation, wales, income deprivation, small area, iterative proportional fitting

**Citation:**

Anderson, B. (2011) *Estimating Small Area Income Deprivation: An Iterative Proportional Fitting Approach*, Centre for Research in Economic Sociology and Innovation (CRESI) Working Paper 2011-02, University of Essex: Colchester.

### About CRESI:

Based in the UK's leading Sociology Department, the [Centre for Research in Economic Sociology and Innovation](#) (CRESI) is the first UK centre for research in economic sociology. With a clear focus on innovation, our research programmes highlight contemporary and historical processes of socio-economic transformation. You can [read about our research](#) and [join our conversation](#).

This work is published under the [Creative Commons Attribution-Non-Commercial-No Derivative Works 2.0 UK: England & Wales License](#)

You are free:



to copy, distribute, display, and perform the work

Under the following conditions:



**Attribution.** You must give the original author credit.



**Non-Commercial.** You may not use this work for commercial purposes.



**No Derivative Works.** You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.
- Nothing in this license impairs or restricts the author's moral rights.

**Your fair dealing and other rights are in no way affected by the above.  
This is a human-readable summary of the [Legal Code \(the full licence\)](#).**

**Table of Contents**

**1 Introduction ..... 5**

**2 Small Area Income Estimation Methods..... 6**

**3 The Iterative Proportion Fitting Approach ..... 7**

**4 Results..... 11**

**5 Validation ..... 14**

**6 Conclusions and future directions..... 17**

**7 Acknowledgements..... 17**

**8 References..... 18**

# 1 INTRODUCTION

As the other chapters in the volume make clear there is increasing demand for the development of small area estimates of a range of socio-economic indicators not only for research and public policy use but also for commercial applications. Sensing this demand, social and economic geographers as well as a number of commercial data providers have developed a range of approaches to the estimation of such variables at small area levels using both econometric (Bates, 2006; Gosh & Rao, 1994; Heady, et al., 2003; Rao, 2003) and spatial microsimulation approaches (Tanton, Mcnamara, Harding, & Morrison, 2009; Williamson, 2005). This work has included a number of attempts to project small area income distributions forwards in time (Ballas, 2004; Dimitris Ballas, et al., 2005) as well as the development of other small area indicators of exclusion and deprivation (Birkin & Clarke, 1989; Williamson & Voas, 2000), ill-health (Ballas, Clarke, Dorling, Rigby, & Wheeler, 2005; Edwards & Clarke, 2009; Mohana, Twigg, Barnard, & Jones, 2005; Morrissey, Clarke, Ballas, Hynes, & O'Donoghue, 2008; D. M. Smith, Harland, & Clarke, 2007), and resource use (Druckman & Jackson, 2008; Williamson, 2001). Other more commercially oriented work has focused on wealth, consumption (expenditure) and lifestyle indicators (Anderson, 2008; Anderson, De Agostini, Laidoudi, Weston, & Zong, 2009; Birkin & Clarke, 2011).

The need for local stakeholders, and especially policymakers to understand the small-area distribution of deprivation has led to a number of approaches to the calculation or estimation of deprivation indicators at levels that cannot be robustly supported by current national surveys. This need has been most clearly met in the United Kingdom by the development of the Indices of Deprivation (IMD) series that now cover the constituent countries of England, Wales, Scotland and Northern Ireland (M Noble, Wright, Smith, & Dibben, 2006).

Income deprivation has always been a key part of the IMD and the income sub-domain provides measures of poverty through aggregation of benefits claimant counts (M Noble, et al., 2008; M Noble, et al., 2004; M Noble, et al., 2006). However with the increasing focus of poverty-related policy on the standard income deprivation measure of the percentage of households below a given income threshold (Eurostat, 2007; Gordon & Townsend, 2000) an alternative approach may be required which can generate income threshold based indicators for inclusion in existing small area deprivation indices (M Noble, et al., 2008).

Unfortunately reliably and robustly estimating household and individual income or income thresholds at small area levels in the United Kingdom is, and has always been, a challenging exercise. Although there have been a number of studies exploring the changing geography of deprivation, there is a general paucity of data, and thus information, relating to the small area geography of household income, wealth and taxation.

The best source of small area socio-economic information in the UK is the census of population but unfortunately the UK census does not provide any information on variables such as household income, wealth and taxation in order to preserve confidentiality and minimize non-response (Marsh, 1993). Aspects of income are of course available from a range of government surveys but in many cases the finest spatial scale to which these survey data are coded are the twelve UK Government Office Regions<sup>1</sup> or, in the case of some surveys the Local Authority District, the main level of UK local administration with mean household population sizes of around 57,000. Even then however, many such surveys provide incomplete geographical coverage sampling only a fraction (and sometimes none) of the residents of any given district.

Overall then, a method aimed at estimating small area income deprivation would need to produce a national small area population distribution, which when aggregated would be within the known error bounds of the official estimates at the regional and national level; support the estimation of threshold based income deprivation indicators; support the small-area level analysis of different policy outcomes; and be estimated at the lowest practical level of spatial geography.

---

<sup>1</sup> See <http://www.statistics.gov.uk/geography/gor.asp>, mean size c 2 million households

In this chapter we briefly review some of the aforementioned approaches to meeting this need before discussing a spatial microsimulation approach based on the iterative proportional fitting algorithm which can deliver against the characteristics listed above. For illustrative purposes we discuss the use of the method in the estimation of small area income deprivation at the Welsh Lower Layer Super Output Area<sup>2</sup> (LSOA) level in order firstly to explore the value of including such estimates in official indicators of multiple deprivation and secondly to enable comparison with and validation against the income domain scores from the Welsh IMD (WIMD) 2005 (M Noble, et al., 2006).

## 2 SMALL AREA INCOME ESTIMATION METHODS

Perhaps the simplest method of estimating small area income levels is the use of variables available in the Census as proxies for income based on correlations between these two variables in national sample surveys. Thus, for example, the mean income for different socio-economic classifications can be calculated from a national income survey such as the United Kingdom's Family Resources Survey (FRS). Given that Census data provides the number of people and/or Household Response Persons within a given socio-economic category, a simple multiplication provides "indirect non-survey designed estimates" of income at geographical levels for which we have socio-economic information. However such simplistic approaches take no account of more complex relationships between socio-economic variables and income distributions nor of the manner in which such relationships may vary in space.

Williamson & Voas (2000) used data from the large-scale UK Census Rehearsal of April 1999 which included a banded income question to test a range of household and area level predictors of small-area income levels (Williamson & Voas, 2000). In this case their analysis suggested that there was a high degree of income heterogeneity even at small area level and that a relatively small number of indicators were the optimal predictors of area level income measures. Williamson suggested that by far the most effective simple proxy for income is the proportion of the economically active population in National Statistics Socio-economic Classification (NS-SEC) categories one and two (Managerial and Professional occupations) and that this finding applies regardless of whether mean income is measured per person, per adult or per number of persons in the household. In addition Williamson's analysis suggested that this indicator performed better than contemporary deprivation indices and interestingly that only 1% of unexplained between-adult income variation could be explained by area level factors such as house prices.

The Office for National Statistics have subsequently developed a regression based approach to the estimation of small area income levels (Bates, 2006; Heady, et al., 2003). Their method involves combining survey data with other data sources that are available on an area basis and is underpinned by the area level relationship between the survey and auxiliary variables such as administrative or census data. In this context they modelled ten variables at the small area level: household income from the Family Resources Survey, household income from the General Household Survey, a measure of social capital, children from ethnic minorities, number of people to help in a crisis, single-parent families, overcrowding, and three measures of poor health. This Small Area Estimation Project (SAEP) methodology has been applied by ONS to produce ward-level estimates of mean household income for 2001/02 and Middle Layer Super Output Area<sup>3</sup> level-level estimates for 2004/5. Unfortunately the approach has not yet been applied to the estimation of threshold-based indicators such as the percentage of households below a given threshold and it also has not been used to create estimates at lower levels of geography such as for LSOAs.

An alternative approach, generally implemented in the commercial market research sector, makes use of 'lifestyle geo-codes' to estimate income distributions down to postcode levels. These estimates generally calculate income distributions using a combination of market research records, postcode level geo-demographic indicators, national census and survey data, and statistical imputation (Webber, 2004). In effect the national income distribution is used as

---

<sup>2</sup> The second level of Census aggregation (containing multiple Census 'Output Areas') containing on average around 600 households.

<sup>3</sup> Administrative areas that are nested within Local Authorities and which are exact aggregates of, on average, 5 LSOAs.

the basis of the income distribution of individual postcodes, but its mean and standard deviation are allowed to vary between postcodes (Williamson and Voas, 2000). However the dependence on lifestyle surveys with potentially unknown response bias, on a small number of lifestyle categories and on the imputation of household characteristics for postcode level address records has been criticized (McLoone, 2002; Williamson & Voas, 2000).

Finally Birkin and Clarke (Birkin & Clarke, 1989; 1988) introduced an approach that sought to combine elements of several of the above to produce not a modelled estimate but synthetic population microdata from which relevant aggregates and summaries could be calculated. This approach emphasised the importance of small area income estimates and was the first study to use what they termed a *spatial microsimulation* method. This method used a combination of Monte-Carlo sampling and iterative proportional fitting to produce small area level microdata for each ward of the City of Leeds.

Since this original work, there have been considerable developments and advances in data availability and computing resources which have enabled experimentation with new techniques that can more easily and efficiently generate more reliable small area microdata. In this context Williamson et al. (Williamson, Birkin, & Rees, 1998) explored different solutions to finding the combination of UK Census household SARs which best fit known small area constraints. They tested various techniques of probabilistic combinatorial optimisation methods such as hill climbing algorithms, simulated annealing approaches and genetic algorithms in order to re-weight cases in the SARs so that a good fit to known Census derived data was achieved when the estimations were re-aggregated to the small area level. Building on these approaches Ballas et al (1999) report the testing of a number of approaches including the deterministic iterative proportional fitting method. Ballas then applied this method to the estimation of small area level trends in equivalised income in York and Leeds between 1991 and 2001 using a combination of Census and British Household Panel (BHPS) data (Ballas, 2004). He concluded that the iterative proportional fitting method was preferable on a number of dimensions including its deterministic nature and relatively efficient algorithm.

More recent work has sought to improve on these initial approaches through the further refinement of methods of error estimation (D. Smith, Clarke, & Harland, 2009) and of the selection of small area constraints (Anderson, 2007; Anderson, De Agostini, & Lawson, 2009; Chin & Harding, 2006; Tanton, et al., 2009).

In summary a range of approaches to the small-area estimation of income have been developed. However very few have attempted to produce the measures of the proportion of households below given income thresholds which are now considered standard in the analysis of income inequalities. In this context the remainder of this chapter presents a method for estimating the percentage of households whose net equivalised income is below 60% of the national median. The method uses the deterministic iterative proportional fitting approach to produce population microdata for each Welsh LSOA using a large sample survey and Census 2001 data for Wales.

### 3 THE ITERATIVE PROPORTION FITTING APPROACH

As we have seen there has been considerable progress in the use of spatial microsimulation to produce small area estimates. The reweighting methodologies briefly introduced above offer considerable potential for the creation of synthetic small area microdata through the re-weighting of national or regional survey microdata, such as the Family Resources Survey (FRS), using data from the Census of population. Put simply the method allocates all households from the sample survey to each small area and then, for each small area, re-weights each household so that the derived small area level tables of aggregate statistics for those re-weighted households match identical tables from the UK Census 2001 (Williamson, et al., 1998). This re-weighting requires the identification of suitable constraint variables that must exist in both the small area (Census) and survey data in identical form (Williamson, et al., 1998). It is these constraints that are the subject of the re-weighting (fitting) process.

As Williamson et al. (1998) and Voas and Williamson (2000) point out there are many ways in which re-weighting methodologies can be fine-tuned through the evaluation of the use of more or different Census small area tables or by changing the model parameters. These design choices can be summarised as choice of re-weighting algorithm; choice of constraints to be used in re-weighting; selection of households from the survey to be used for each small area

and whether or not to require integer weights (i.e. produce "whole households").

A wide range of techniques have been proposed for the re-weighting of cases ranging from iterative proportional fitting through simulated annealing to linear programming and complex combinatorial optimization methods (Ballas & Clarke, 2001; Ballas, Clarke, & Turton, 1999; Williamson, et al., 1998). Birkin and Clarke (Birkin & Clarke, 1989; Birkin & Clarke, 1988) demonstrate how Iterative Proportional Fitting (IPF) and Monte Carlo sampling can be employed to generate a wide range of attributes at the small area level. The IPF method is well-established and appears in a multitude of guises, from balancing factors in spatial interaction modeling through to the RAS method in economic accounting and its behaviour is relatively well known (Birkin & Clarke, 1988; Simpson & Tranmer, 2005; Wong, 1992).

In essence the method we have developed allocated all (or a specific selection of) households from the FRS to each Welsh LSOA and then iteratively re-weighted each case using the iterative proportional fitting algorithm so that LSOA level tables of aggregate statistics matched identical tables from the UK Census 2001.

### 3.1 Definition of Income

The income survey data used was the Welsh subsample of the FRS 2003/4 and FRS 2004/5 and the base income variable used was the sum of all net household income from:

- Earnings & self employment (net of income tax and national insurance payments)
- Investments
- Disability benefits
- Retirement pensions plus any income support or pension credit
- Working Tax Credit and/or Child Tax Credit received
- Other pensions
- Other benefits
- Other/remaining sources

In order to align the income values with the official UK Department for Work and Pensions' Households Below Average Income (HBAI) definitions (DWP, 2007: Appendix 1) the following expenditures were then removed to produce the net income before housing costs (BHC):

- Domestic rates / council tax;
- Contributions to occupational pension schemes (including all additional voluntary contributions (AVCs) to occupational pension schemes, and any contributions to stakeholder and personal pensions);
- Insurance premia payments made in case of sudden loss of earnings;
- All maintenance and child support payments, which are deducted from the income of the person making the payment;
- Parental contributions to students living away from home;
- Student loan repayments.

To calculate after housing costs (AHC) income '*the total amount spent on water and sewerage rates, rent, mortgage interest, household rent, structural insurance (adjusted for combined cases to be consistent with HBAI) and service charges*' (DWP, 2007: Appendix 1) was removed from the before housing costs income variable.

In common with the official HBAI definitions, the Organisation for Economic Co-operation and Development's modified equivalisation scale<sup>4</sup> was then used to control for household composition and to produce an equivalised measure of household income before and after

---

<sup>4</sup> A scale that controls for household size/composition. Equivalised income = income/w where  $w = 0.67a + 0.33b + 0.2c + 0.33d$  and a = first adult, b = number of second (e.g. spouse) and other adults, c = number of children aged < 14 and d = number of children aged 14 or more.



housing costs. These were then used as the basis for the calculations of the Welsh BHC/AHC medians and thence the allocation of households to the two indicator groups (above or below 60% of the relevant Welsh median) to create a within-Wales poverty indicator.

**Table 1: FRS BHC/AHC Households Below Average Income results for Wales (weighted sample survey)**

	BHC	AHC
2003/4		
N	1278	1278
% HBAI	12.51%	17.76%
2004/5		
N	1239	1239
% HBAI	13.48%	16.87%
2003/5		
N	2517	2517
% HBAI	12.99%	17.32%

It should be noted that households with negative income were retained. Households reporting negative BHC income constitute 0.55% of Welsh households in 2003-4 (0.81% in 2004/5) whilst 0.86% (1.37%) report negative AHC income. It was not expected that retaining negative income would therefore have any significant effect on the indicators as they will not substantially effect the median derived calculations.

### 3.2 Choice of constraint variables

Having already determined that IPF was to be used to re-weight the cases to fit the small area tables it was necessary to identify the constraint variables on which the IPF process would operate. The set of constraint variables must be:

1. Common to both the FRS and the Census or at least can be derived from them;
2. Available at the household level – as the poverty indicator to be estimated is at the household level;
3. Known to be reasonable predictors of the indicator at the small area level
4. Reasonably good predictors of the indicator at the micro (household) level.

A review of Census and FRS data was used to produce a list of variables that satisfied criteria one and two and recommendations from the literature (Williamson, 2005; Williamson & Voas, 2000) were used to filter these variables according to criteria three to produce a list of candidate constraints (Anderson, 2009).

Finally criteria four was tested within the FRS using standard logistic regression techniques to model the relationship between the micro level constraints and the probability of a household having a net equivalised income below 60% of the national median. The r square value was used as an indicator of the value of the constraint variables but in contrast to previous work which reported the use of repeated bi-variate regressions to test each variable independently (Chin & Harding, 2006), a stepwise multivariate method was used. The multivariate approach meant that correlations between constraint variables were taken into account and thus the 'pure' effects of each constraint was revealed whilst the use of the stepwise technique automatically included only those variables which had a statistically significant effect on the model and ordered the resulting indicators in decreasing order of the magnitude of their affects.

The overall model pseudo r-squared score was then used as an indicator of how well the included constraints predicted the outcome variables (in this case the BHC or AHC HBAI) at the household level and thus an indicator of the confidence we can have in the robustness of the eventual results. In addition, because the IPF technique iteratively re-weights a series of constraints, the last constraint is necessarily fitted perfectly. It is therefore important that the constraints are used in an order that represents their increasing predictive power so that the 'best' constraint is fitted last and the stepwise results allowed us to establish this ordering.

Table 2 summarises these results and shows that we can be justified in pooling the 2003-4 and

2004-5 FRS data since the predictors of each indicator at the household level were essentially identical although it is interesting to note that with the larger pooled sample (03-05) there were additional significant constraint variables - HRP gender in the case of BHC and HRP age in the case of AHC.

**Table 2: Significant constraints (in decreasing order of explanatory power, see (Anderson, 2009) for full results)**

	2003-4		2004-5		2003-05 pooled	
	BHC	AHC	BHC	AHC	BHC	AHC
	Employment Status	Employment Status	Employment Status	Employment Status	Employment Status	Employment Status
	Number of Earners	Tenure	Number of Earners	Number of Earners	Number of Earners	Tenure
	Tenure	Number of Earners		Tenure	Tenure	Number of Earners
					Gender of HRP	HRP age
Pseudo r square	0.122	0.234	0.126	0.219	0.129	0.232

### 3.3 Small area IPF algorithm implementation

As previously discussed these constraints were then used at the small area (LSOA) level to iteratively re-weight the FRS to fit each Welsh LSOA and so produce an estimate of the %HBAI for each LSOA for each indicator. Whilst results for 2003-4 and 2004-5 were generated separately we report only those for the pooled 2003-4-5 data using the constraints identified above.

Following Ballas et al (2005) we implemented a regional weighting scheme so that only households belonging to the same region as the particular LSOA are allocated to it. Ballas et al also report using a process of integerisation to select the 'best fit' n weighted households for a given area where n was the number of households required for the ward. Ballas et al report that this integerisation produced some extremely poor results when tested against the census distributions and described a swapping algorithm to swap households between their 1991 wards in order to reduce errors and produce a better fit.

Since it is inevitable that the integerisation process will reduce within-zone variation and for our purposes it was not necessary that each small area was allocated a whole number of households, we did not implement the integerisation process. Instead our simplified method allowed the final household weights for each small area to remain fractional so that all possible survey households were retained. In our experience this simplified method produced distributions that performed at least as well as Ballas et al's more complex combination of integerisation and household swapping.

The objective was to produce a set of weights linking all households from the relevant Government Office Region to all LSOAs in that region in the sense that the weights represent the 'fractional existence' of the each household in each LSOA. Conceptually the results can be thought of as a matrix of LSOAs (rows) and households (columns) where each cell contains the weight for a given household in a given LSOA.

To do this two sets of tables were required for each constraint for each LSOA: the Census 2001 small area tables for the constraints (e.g. Table 3) and the analogous small area tables constructed from the FRS households for the region in which the zone was found (e.g. Table 4)

**Table 3: Small Area Table for number of earners derived from Census 2001 for the first LSOA in Wales**

Zonocode	Number of households	Number of earners = 0	Number of earners = 1	Number of earners = 2	Number of earners = 3+
W01000001	517	294	132	85	6

**Table 4: 'Small Area Table' for number of earners derived from the FRS 2003/4/5 for Wales**

Number of households	Number of earners = 0	Number of earners = 1	Number of earners = 2	Number of earners = 3+
1308	608	333	320	47

Starting with LSOA 1 all household weights ( $w_i$ ) were initially set to 1 whilst the weights of households that did not belong to the same region as the area in question were set to 0 rather than  $w_i$  to implement the regional weighting scheme.

**Table 5: First four Zone 1 households with initial weights**

case	region	age	rooms	persons	nssec	comp	nearners	$W_i$
26115	10	2	3	0	1	2	1	1
26116	10	2	2	0	3	2	0	1
26117	10	2	3	4	0	0	2	1
26118	10	4	3	0	0	2	1	1
..	..	..	..	..	..	..	..	1

Then for each constraint in turn, the weights were adjusted using the formula:

$$NW_h = w_{ih} * c_{hj} / S_{hj}$$

where  $NW_h$  was the new household weight for household  $h$ ,  $w_{ih}$  was the initial weight for household  $h$ ,  $c_{hj}$  was element  $hj$  of the census data table (Table 3) and  $s_{hj}$  was element  $hj$  of the FRS statistical table (Table 4).

As an example, using the number of earners constraint Table 6 shows the calculations of the first weights for the first four households so that the FRS sample fits the Census distributions on this one dimension.

**Table 6: First four Zone 1 households with weights after fitting to constraint 1**

case	region	nearners	$W_1$
26115	10	1	= 1 * (132/333) = 0.396
26116	10	0	= 1 * (294/608) = 0.484
26117	10	2	= 1 * (85/320) = 0.266
26118	10	1	= 1 * (132/333) = 0.396
..	..	..	..

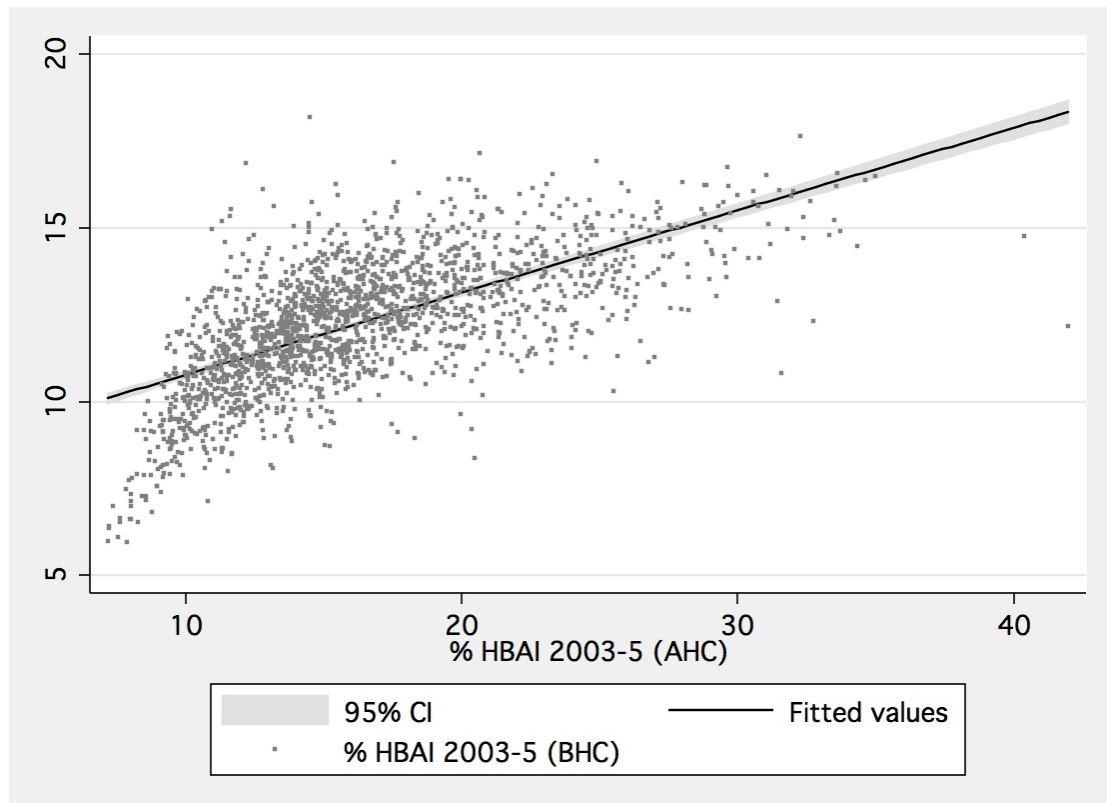
Having adjusted the weights for the first constraint the process then moved sequentially through each constraint variable multiplying each new weight by that produced by the previous step. Since the last constraint to be fitted will necessarily be fitted perfectly, it was necessary to order the variables in 'r sq contribution' order (cf. Table 2) so that the last to be fitted was the one which accounted for the most variation in the outcome variable of interest (% HBAI in this case).

Having passed over all constraints once, the process then looped back to constraint one and repeated the re-weighting starting from the weight produced in the last step (by the last constraint). Ballas et al found that iterating the procedure between 5 and 10 times produced weights that reduced the error in fitting households to areas to a point where it no longer declined. Our experimentation suggested that 10 iterations were sufficient to achieve a stable indicator value. (Anderson, 2009) Thus after iterating over the re-weighting procedure 10 times the simulation then moved on to the next zone and repeated the process.

Calculating the % HBAI was thus a straightforward matter of summing the weighted indicator for each area and dividing by the number of households in that area.

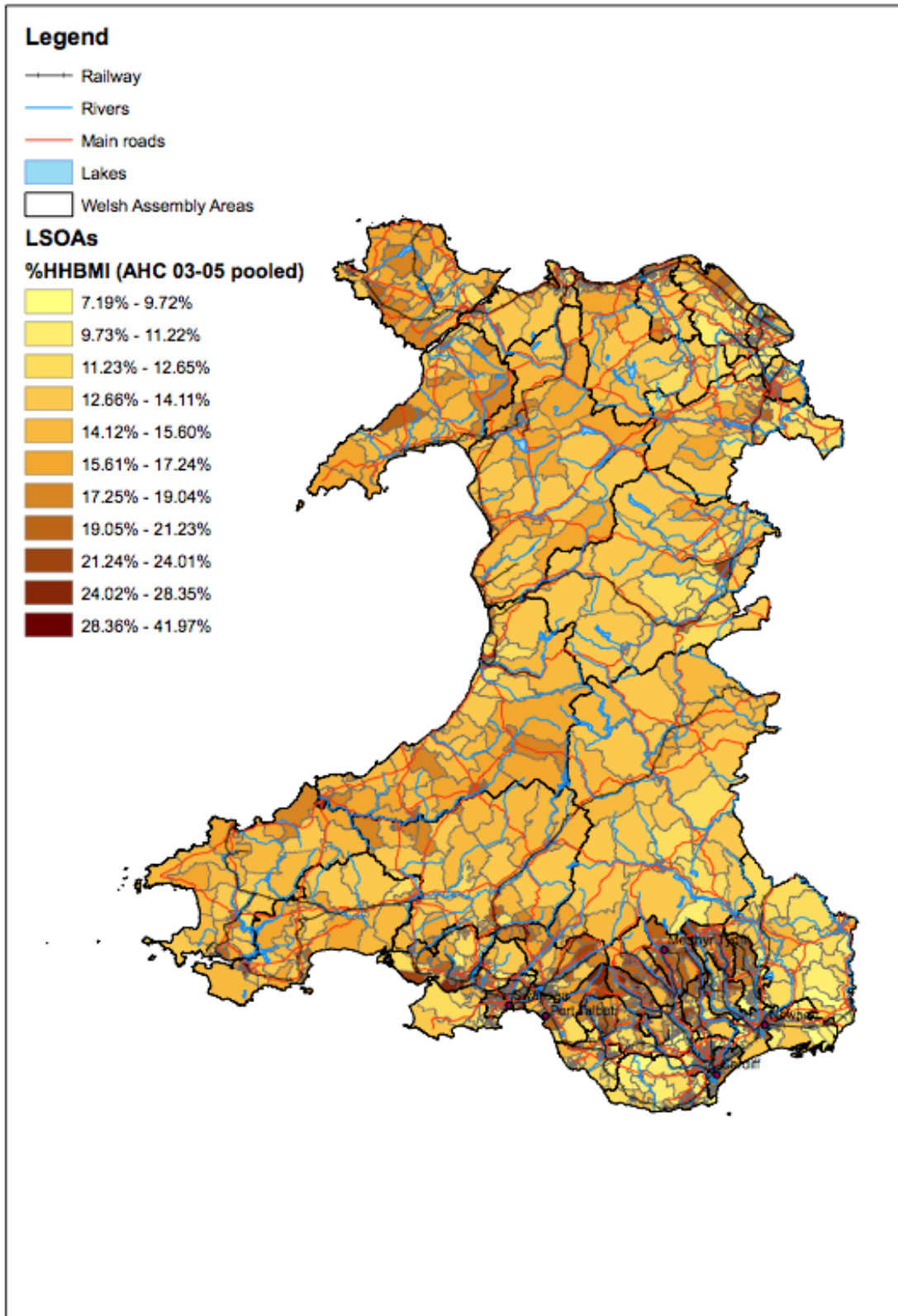
## 4 RESULTS

Figure 1 compares estimated before housing costs (BHC) and after housing costs (AHC) % HBAI for LSOAs in Wales. As we can see the two indicators produced slightly different results at the upper end of the distribution (higher income deprivation) although rather similar results for the least deprived and this was confirmed by a Spearman rank correlation of only 0.679.



**Figure 1: Comparison of BHC and AHC HBAI indicators at LSOA for Wales (Census 2001, FRS 2003-05 pooled)**

The spatial distributions of the BHC indicator (not shown) suggested a higher concentration of poor households in the former mining areas of South Wales and in the coastal areas on the Pembrokeshire/Ceredigion border as well as in other pockets in specific urban areas. In contrast the areas with the highest % HHBMI according to the AHC indicator (Figure 2) were more markedly concentrated in the Valleys and South Wales urban areas.

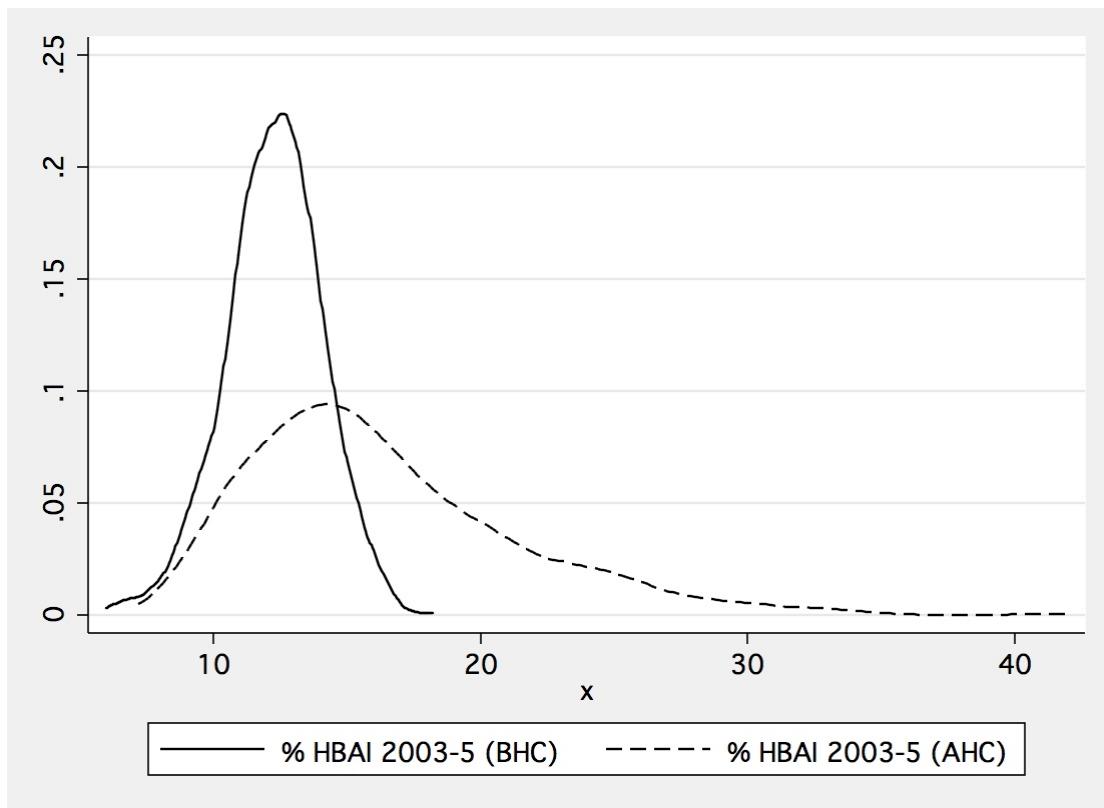


**Figure 2: Spatial distribution of %HBAI in Wales (AHC, LSOA level, Census 2001, FRS 2003-05 pooled)**

Overall however the AHC indicator was considerably more diffuse in its distribution and thus may be a better 'relative poverty' indicator in comparison to the rather tighter BHC distribution

which supports less differentiation (Figure 3).

It should therefore be apparent that the utility of each indicator in a revised Index of Multiple Deprivation will depend on the political objective and outcome desired since they reveal slightly different patterns of poverty.



**Figure 3: Kdensity distributions of BHC and AHC % HBAI indicators at LSOA level for Wales (Census 2001, FRS 2003-05 pooled)**

## 5 VALIDATION

In order to test the validity of the estimated distributions of the % HBAI we made three kinds of comparisons:

- Comparison of the estimated results with initial source survey results (i.e. at regional or country level) to check internal validity and that the process accurately re-created inter-regional or inter-country variation. In this case we used the FRS;
- Comparison of the estimated constraint counts with initial census-derived constraint counts to check internal validity. This was the analysis of Total Absolute Error (TAE) discussed in Ballas et al (Ballas, et al., 1999; Ballas, Dorling, Anderson, & Stoneman, 2006; D. Smith, et al., 2009);
- Comparison of estimated LSOA level results with other known small area estimates. In this case no equivalent small area data were available but instead we compared the results with the income domain score of the Welsh IMD 2005;

Table 7 shows the % HBAI indicators (and 95% confidence interval) as calculated from the relevant source FRS data and as estimated from the spatial microsimulation process. Overall there appeared to be a tendency to slightly underestimate % HBAI compared to the source FRS results. In general however we would expect the microsimulation result to lie within the 95% confidence interval of the survey estimate and as can be seen, the spatially microsimulated estimates provided a reasonable fit since they lay within these boundaries.

**Table 7: Comparison of simulated mean % HBAI results with source (FRS 2003-4-5 pooled) results.**

	Source FRS 2003-4-5 pooled	SE mean	95% CI (+/-)	Spatial microsimulation (FRS 2003-4-5 pooled/Census 2001)	Difference
% HBAI (BHC)	12.992%	0.670	1.314	12.259%	-0.733%
% HBAI (AHC)	17.322%	0.754	1.479	16.294%	-1.028%

Turning to the constraint count error analysis, by entering the constraint counts as variables to be estimated it is possible to compare the initial 'true' census constraint household counts with the estimated counts following the spatial microsimulation procedure. The total absolute error (TAE) is the difference for each constraint category for each area summed over all areas whilst the standardised absolute error (SAE) is TAE divided by the total number of units (in this case households).

Whilst minimising the difference between the 'true' and estimated counts is the objective it is not yet clear in the literature what values of error are acceptable although Smith et al suggest that an SAE of less than 20% and ideally less than 10% in 90% of the areas is desirable especially where the prevalence rate of the phenomenon of interest is low (D. Smith, et al., 2009).

The % HBAI models for Wales performed substantially better than this and elsewhere we disaggregated the SAE to reveal the constraints that showed the poorest fit (Anderson, 2009). The analysis suggested that levels of error were relatively low for both indicators with the largest error being for households with no earners (11.9% in each case), the mean being no larger than 2.1% for any constraint and in all cases 90% of areas had SAE rates of less than 4%. As expected we also confirmed that the order of the constraints means that the last category to be fitted (HRP Employment status) does so perfectly.

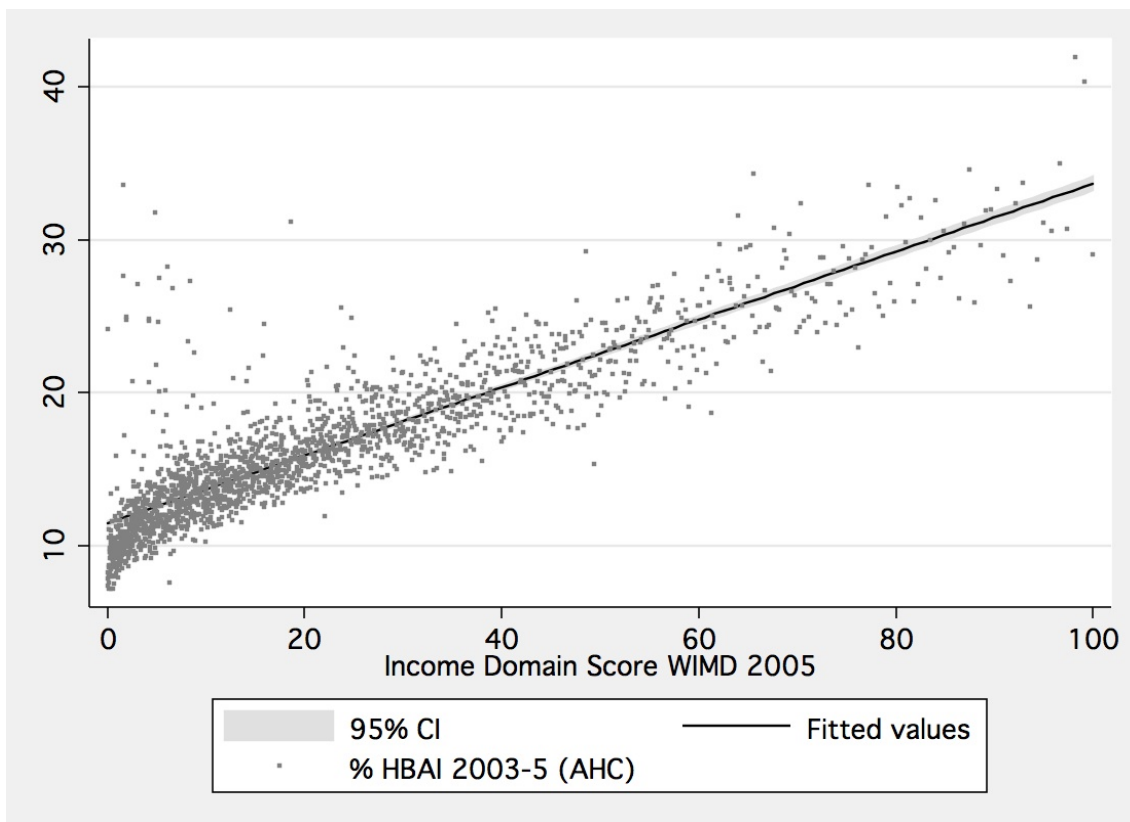
Finally Figure 4 and Figure 5 show the fit between the Welsh IMD 2005 income domain score and the simulated % HBAI indicators at LSOA level. As we would expect there was a strong rank order correlation between % HBAI using the equalised indicator and the WIMD income domain score and this was especially the case for the after housing costs (AHC) indicator.

Figure 5 suggests the presence of a number of outliers which were low on the WIMD 2005 income score but relatively high on the simulated % HBAI. Deeper analysis reveals that the two LSOAs which were in both the top 10% of HBAI (AHC) and the bottom 10% of the WIMD 2005 income score<sup>5</sup> had high or relatively high proportions of students (who could not claim relevant benefits) according to the 2001 Census. This suggested that one of the main differences between the Indices of Deprivation income domain results and the HBAI (AHC) results was the inclusion of low income student households. LSOAs with higher proportions of students were therefore likely to appear to be 'more deprived' using the HBAI indicator than would be the case for the WIMD income domain score.

<sup>5</sup> Somewhat counter intuitively this means their WIMD 2005 income score would be low (i.e. not deprived).



**Figure 4: WIMD 2005 income domain score vs spatial microsimulation results for BHC indicator (Spearman rho = 0.6041)**



**Figure 5: WIMD 2005 income domain score vs spatial microsimulation results for AHC indicator (Spearman rho = 0.8834)**



## 6 CONCLUSIONS AND FUTURE DIRECTIONS

Overall the results of this preliminary work on estimating the proportion of households below HBAI at the small area level were encouraging. The results provided a synthetic household dataset which reproduced the Welsh % HBAI (BHC/AHC) as measured by the UK Family Resources Survey and which also produced a good fit to the Welsh IMD 2005 income domain score at the small area level. This was especially true for the AHC measure.

The results also suggested that a focus on % HBAI, and especially on the AHC indicator would present slightly different spatial distributions of income deprivation than would the WIMD 2005 income domain score. In particular there would be differences where students make up a high proportion of household response persons. This is of course likely to effect specific urban rather than rural areas. In addition the differences between the results for the BHC and AHC indicators mean that consideration needs to be given to which would be the 'best' one to use in a future revision of the Welsh IMD. This cannot be answered by this chapter as it is dependent on the policy context and the uses to which the Index and its components will be put.

The analysis of errors (SAE) suggested that in some small areas the spatial microsimulation method produced a less effective fit than in others for some constraints. This may have been because these areas were made up of an unusual combination of household types and future work could investigate extending the spatial microsimulation method to account for such areas and thus to reduce overall error still further.

The Iterative Proportional Fitting method itself performed in a robust, deterministic manner in this context. This determinism meant that variations in input data coding, constraint ordering or small area table recoding were the only source of variation in the small area estimates. This proved extremely useful because it allowed the testing of different combinations of constraints and data coding options without the additional uncertainty caused by a probabilistic re-weighting method.

As we have discussed we also assumed that the order of the re-weighting iterations mattered since the last constraint always fitted perfectly. We assumed that this then necessitated the use of the nested regression analysis to determine the 'best' constraint variable order. However there is as yet no evidence that this 'best' order produces a substantially better fit than, for example, a random ordering and we anticipate testing this assumption in future work.

Finally we would also suggest that the ability of the IPF algorithm to produce fractional weights proved crucial to the reconstruction of accurate aggregated estimates for comparison with the original survey data (c.f. Table 7). It also enabled us to retain all relevant households in our synthetic small area samples and so increased their (weighted) heterogeneity. This would prove crucial if these data were then to be used as a basis for the microsimulation modeling and thus small area analysis of, for example, policy intervention scenarios.

## 7 ACKNOWLEDGEMENTS

This work is based on data provided through EDINA UKBORDERS with the support of the ESRC and JISC and uses boundary material which is copyright of the Crown and the Post Office.

Census data was originally created and funded by the Office for National Statistics and are distributed by the Census Dissemination Unit, MIMAS (University of Manchester). Census Output is Crown copyright and are reproduced with the permission of the Controller of HMSO.

The FRS is collected and sponsored by the Department for Work and Pensions and distributed by the UK Data Archive, University of Essex, Colchester. FRS data is copyright and is reproduced with the permission of the Controller of HMSO and the Queen's Printer for Scotland.

The WIMD 2004 were constructed by the Social Disadvantage Research Centre at the Department of Social Policy and Social Research at the University of Oxford and distributed by the Welsh Assembly Government.

We would also like to thank Professor Holly Sutherland (ISER, University of Essex) for advice on the treatment of negative income and housing costs.

This work was sponsored by the Welsh Assembly Government.

## 8 REFERENCES

- Anderson, B. (2007). *Cash in, Cash out: Spatially microsimulating household income and expenditure at small area levels*. Paper presented at the Royal Statistical Society Conference 2007.
- Anderson, B. (2008). *Time to play: combining time-use surveys and census data to estimate small area distributions of potentially ICT mediated leisure*. Paper presented at the AoIR 8.
- Anderson, B. (2009). *Welsh Small Area Estimates of Income Deprivation*. Colchester: University of Essex.
- Anderson, B., De Agostini, P., Laidoudi, S., Weston, A., & Zong, P. (2009). Time and Money in Space: Estimating Household Expenditure and Time Use at the Small Area Level in Great Britain. In A. Zaidi, A. Harding & P. Williamson (Eds.), *New Frontiers in Microsimulation Modelling. Public Policy and Social Welfare Vol. 36*. Aldershot: Ashgate.
- Anderson, B., De Agostini, P., & Lawson, T. (2009). *Estimating income, expenditure and time-use within small areas*. Paper presented at the ESRC Microsimulation Seminar Series Workshop III 'Moving beyond tax-benefit and demographic modelling'.
- Ballas, D. (2004). Simulating trends in poverty and income inequality on the basis of 1991 and 2001 census data: a tale of two cities. *Area*, 36(2), 146-163.
- Ballas, D., & Clarke, G. (2001). Modelling the local impacts of national social policies: a spatial microsimulation approach. *Environment and Planning C: Government and Policy*, 19, 587-606.
- Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B., & Rossiter, D. (2005). SimBritain: A Spatial Microsimulation Approach to Population Dynamics. *Population, Space and Place*, 11, 13-34.
- Ballas, D., Clarke, G., Dorling, D., Rigby, J., & Wheeler, B. (2005). *Using Geographical Information Systems and spatial microsimulation for the analysis of health inequalities*. Paper presented at the 10th International Symposium on Health Information Management Research - iSHIMR 2005.
- Ballas, D., Clarke, G., & Turton, I. (1999, 25-28 July, 1999). *Exploring Microsimulation Methodologies for the Estimation of Household Attributes*. Paper presented at the 4th International Conference on GeoComputation, Mary Washington College, Virginia, USA.
- Ballas, D., Dorling, D., Anderson, B., & Stoneman, P. (2006). *Assessing the feasibility of producing small area income estimates: Phase I project report*. Sheffield: Department of Geography, University of Sheffield.
- Bates, A. (2006). Methodology used for producing ONS's Small Area Population Estimates. *Population Trends*, 125, 30-36.
- Birkin, M., & Clarke, G. (1989). The generation of individual and household incomes at the small area level using Synthesis. *Regional Studies*, 23(6), 535-548.
- Birkin, M., & Clarke, M. (1988). SYNTHESIS – a synthetic spatial information system for urban and regional analysis: methods and examples. *Environment and Planning A*, 20, 1645-1671.
- Birkin, M., & Clarke, M. (2011). Spatial Microsimulation Models: A Review and a Glimpse into the Future. In J. Stillwell & M. Clarke (Eds.), *Population Dynamics and Projection Methods*. London: Springer.
- Chin, S.-F., & Harding, A. (2006). *Regional Dimensions: Creating Synthetic Small-area Microdata and Spatial Microsimulation Models*. Canberra: National Centre for Social and Economic Modelling, University of Canberra.
- Druckman, A., & Jackson, T. (2008). Household energy consumption in the UK: A highly geographically and socio-economically disaggregated model. *Energy Policy*, 36(8), 3177-3192.
- DWP (2007). *Households Below Average Income (HBAI) 1994/95-2005/06*. London: Department of Work and Pensions.
- Edwards, K. L., & Clarke, G. P. (2009). The design and validation of a spatial microsimulation

model of obesogenic environments for children in Leeds, UK: SimObesity. *Social Science & Medicine*, 69(7), 1127-1134.

- Eurostat (2007). Inequality of income distribution (S80/S20 income quintile share ratio), at-risk-of-poverty rate and at-persistent-risk-of-poverty rate. *Key indicators on EU policy – Structural indicators – Social Cohesion – Living conditions* Retrieved July 4th, 2008, from [http://europa.eu.int/estatref/info/sdds/en/strind/socohe\\_di\\_base.htm](http://europa.eu.int/estatref/info/sdds/en/strind/socohe_di_base.htm)
- Gordon, D., & Townsend, P. (2000). *Breadline Europe: The Measurement of Poverty*. Bristol: Policy Press.
- Gosh, & Rao, J. K. (1994). Small Area Estimation: An Appraisal. *Statistical Science*, 9(1), 55-76.
- Heady, P., Clarke, P., Brown, G., Ellis, K., Heasman, D., Hennell, S., et al. (2003). *Model-Based Small Area Estimation Series No. 2*. London: Office for National Statistics.
- Marsh, C. (1993). Privacy, confidentiality and anonymity in the 1991 Census. In A. Dale & C. Marsh (Eds.), *The 1991 Census User's Guide* (pp. 111-128). London: Her Majesty's Stationary Office.
- McLoone, P. (2002). *Commercial Income Data: Associations with Health and Census Measures*.
- Mohana, J., Twigg, L., Barnard, S., & Jones, K. (2005). Social capital, geography and health: a small-area analysis for England *Social Science & Medicine*, 60, 1267–1283
- Morrissey, K., Clarke, G., Ballas, D., Hynes, S., & O'Donoghue, C. (2008). Examining access to GP services in rural Ireland using microsimulation analysis. *Area*, 40(3), 354-364.
- Noble, M., McLennan, D., Wilkinson, K., Whitworth, A., Barnes, H., & Dibben, C. (2008). *Indices of Deprivation 2007*. London: Communities and Local Government.
- Noble, M., Wright, G., Dibben, C., Smith, G., McLennan, D., Anttila, C., et al. (2004). *Indices of Deprivation 2004*. London: Office of the Deputy Prime Minister.
- Noble, M., Wright, G., Smith, G., & Dibben, C. (2006). Measuring multiple deprivation at the small-area level. *Environment and Planning A*, 38(1), 169-185.
- Rao, J. K. (2003). *Small Area Estimation*. London: Wiley.
- Simpson, L., & Tranmer, M. (2005). Combining sample and census data in small area estimates: Iterative Proportional Fitting with standard software. *The Professional Geographer*, 57(2), 222-234.
- Smith, D., Clarke, G., & Harland, K. (2009). Improving the synthetic data generation process in spatial microsimulation models. *Environment and Planning A*, 41, 1251 - 1268.
- Smith, D. M., Harland, K., & Clarke, G. (2007). *SimHealth: Estimating Small Area Populations using Deterministic Spatial Microsimulation in Leeds and Bradford*. Leeds: University of Leeds.
- Tanton, R., Mcnamara, J., Harding, A., & Morrison, T. (2009). Small Area Poverty Estimates for Australia's Eastern Seaboard in 2006. In A. Zaidi, A. Harding & P. Williamson (Eds.), *New Frontiers in Microsimulation Modelling. Public Policy and Social Welfare Vol. 36*. Aldershot: Ashgate.
- Webber, R. (2004). *The Relative Power of Geodemographics vis a vis person and household level demographic variables as discriminators of consumer behaviour*. London: UCL.
- Williamson, P. (2001). An Applied Microsimulation Model: Exploring Alternative Domestic Water Consumption Scenarios. In G. Clarke & M. Madden (Eds.), *Regional Science in Business*. London: Springer.
- Williamson, P. (2005). *Income imputation for small areas*. Liverpool: University of Liverpool.
- Williamson, P., Birkin, M., & Rees, P. (1998). The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning A*, 30, 785-816.
- Williamson, P., & Voas, D. (2000). Income Estimates for Small Areas: Lessons from the Census Rehearsal. *BURISA*, 146, 2-10.
- Wong, D. (1992). The Reliability of Using the Iterative Proportional Fitting Procedure. *Professional Geographer*, 44(3), 340-348.

