
Article:
Whitworth, A., Carter, E., Ballas, D. et al. (1 more author) (2017) Estimating uncertainty in spatial microsimulation approaches to small area estimation: a new approach to solving an old problem. Computers, Environment and Urban Systems, 63. pp. 50-57. ISSN 0198-9715

https://doi.org/10.1016/j.compenvurbsys.2016.06.004

This is the open access pre-publication version of the article published in Computers, Environment and Urban Systems (2016). Please cite the CEUS published version.

Estimating uncertainty in spatial microsimulation approaches to small area estimation: a new approach to solving an old problem

Whitworth, A[1], Carter, E, Moon, G and Ballas, D

Abstract

A wide range of user groups from policy makers to media commentators demand ever more spatially detailed information yet the desired data are often not available at fine spatial scales. Increasingly, small area estimation (SAE) techniques are called upon to fill in these informational gaps by downscaling survey outcome variables of interest based on the relationships seen with key covariate data. In the process SAE techniques both rely extensively on small area Census data to enable their estimation and offer potential future substitute data sources in the event of Census data becoming unavailable. Whilst statistical approaches to SAE routinely incorporate intervals of uncertainty around central point estimates in order to indicate their likely accuracy, the continued absence of such intervals from spatial microsimulation SAE approaches severely limits their utility and arguably represents their key methodological weakness. The present article presents an innovative approach to resolving this key methodological gap based on the estimation of variance of the between-area error term from a multilevel regression specification of the constraint selection for iterative proportional fitting (IPF). The performance of the estimated credible intervals are validated against known Census data at the target small area and show an extremely high level of performance. As well as offering an innovative solution to this long-standing methodological problem, it is hoped more broadly that the research will stimulate the spatial microsimulation community to adopt and build on these foundations so that we can collectively move to a position where intervals of uncertainty are delivered routinely around spatial microsimulation small area point estimates.

Keywords: small area estimation; spatial microsimulation; iterative proportional fitting; credible intervals; confidence intervals; variance estimation.

1. Introduction

A wide range of user groups from policy makers to media commentators desire ever more spatially detailed information in order to better understand their communities, better target resources and better plan activities and interventions. Census data are the obvious key data source here but although in many countries the availability of census and administrative data with high spatial resolution has increased dramatically in recent years key variables of interest frequently remain impossible to access at small area resolutions or with sufficient regularity to capture change over time.

In response to this need, small area estimation (SAE) methodologies – have become increasingly used and demanded as an important means of providing spatially detailed

[1] Corresponding author: Dr Adam Whitworth, Dept of Geography, Univ Sheffield, UK. adam.whitworth@sheffield.ac.uk

insights. These methodologies typically use survey data and with such data direct estimates of small area measures are rarely possible as survey respondents are seldom available from all small areas within a wider target setting. Instead, researchers have methodologies developed regression-based and spatial microsimulation approaches. These have given insights that would not otherwise be possible (e.g. income, fear of crime, healthy behaviours to name but a few UK examples of non-Census variables that are of spatial interest to policy makers)(Marshall 2012; Whitworth, 2013).

Despite this growing interest, one of the two chief methodological approaches to SAE – the family of spatial microsimulation methods – is at present undermined by its key inability to deliver intervals of uncertainty around its central point estimates. This is a critical requirement of any SAE method (Rao, 2005; Chatterjee et al., 2008) and *the* key (and significant) weakness of spatial microsimulation approaches (Tanton et al., 2014; Nagle et al., 2014). Regression-based SAE approaches do not suffer from this methodological Achilles' heel and hence make a strong claim at present to be the preferred approach, yet this is to overlook the possible advantages that spatial microsimulation methods have the potential to deliver if they could be developed to also be able to also estimate intervals around their central point estimates. It is this current inability to estimate credible intervals around point estimates within spatial microsimulation approaches to SAE that therefore motivates this paper to offer an innovative proposed solution to this key weakness.

2. Methodological approaches to small area estimation

As summarised elsewhere (Bishop et al., 1975; Ghosh and Rao, 1994; Rao, 2003; Rahman 2008; Marshall 2012; Whitworth, 2013), various SAE methodologies currently exist and can broadly be described as falling within the two broad churches of spatial microsimulation techniques and statistical regression-based techniques, with further alternative variants and implementations within each broad approach.

Statistical SAE follows logically from the basic notions of model-based prediction and imputation. A statistical model is developed using survey data and its coefficients are then applied to data that match the model explanatory variables but are available for all small areas of interest. A variety of alternative model specifications can be used, with the choice of modelling specification depending on the degree of complexity sought, the nature of the variable to be estimated, the type of estimates desired (e.g. mean, median, or distributional values), the nature of small area covariate data able to be sourced, and the level and structure of the data (Ghosh and Rao, 1994; Rao, 2003; Chambers and Tzavidis, 2006; Tzavidis et al., 2010; Pfeffermann, 2013). Whichever statistical technique is used, the result is a set of small area estimates accompanied by intervals around those central point estimates in order to give an indication of their likely plausible range.

Within the family of spatial microsimulation techniques three alternative methodologies dominate the literature – iterative proportional fitting (IPF), combinatorial optimisation (CO) and generalised regression reweighting (GREGWT). These approaches have been applied to diverse small area research projects in a wide range of national contexts (Voas and Williamson, 2000; Ballas et al 2006; Anderson 2007; Rahman et al. 2010; Tanton et al. 2011;

Birkin and Clarke 2011; Hermes and Poulson, 2012; Tanton and Edwards, 2013). The three approaches seek in differing ways to 'fit' the survey cases as closely as possible to the multi-dimensional characteristics of each separate small area for the set of selected key explanatory variables (termed 'small area constraints' in the literature) for which aggregate small area totals are known, in effect using the survey data to create synthetic micro-populations for each target small area in turn and then using this to pick off estimates of the outcome variable of interest.

The way that the three microsimulation methods achieve their goal differs in important respects. CO operates by selecting the required number of individuals or households from the survey data for the target small area in question. These survey cases are then swapped with cases not yet selected in an attempt to optimise the fit between the cases selected and the characteristics of the small area, with different possible algorithms used to assess whether the swaps have resulted in an improvement to the fit. In contrast, IPF and GREGWT reweight all survey cases to the constraint characteristics for each small area such that, taken together, the survey cases optimally match each small area's profile across the selected constraint variables. This position is reached when the reweighting process stabilises and no longer adjusts the weights. At this point no further improvements in the fit of the constraints between the survey cases and the target small area profile on those constraints is possible and the method is said to have converged. In an IPF approach this reweighting of the survey cases occurs sequentially across the constraint variables in turn. Whichever of these three spatial microsimulation methods is used, however, the result is a set of small area point estimates that can be readily calculated from the outcome values across either the reweighted (IPF and GREGWT) or selected (CO) survey cases for that target small area.

In many ways, therefore, spatial microsimulation and statistical approaches to SAE offer alternative methodological routes to the same desired end point of a set of small area estimates of an outcome of interest that would not otherwise be available. However, one (quite literally) significant way in which the two broad approaches to SAE differ is in terms of the delivery of bounds of expected precision around the central small area point estimates. For statisticians the creation of confidence intervals around point estimates is deeply engrained into thinking and work practices and intervals around statistically derived small area point estimates are produced as a matter of course. These help users to understand the likely precision of the resulting small area estimates and, in doing so, to help users to consider the weight and confidence that they may wish to place in the estimates. For policy makers this is particularly important given their frequent need to use small area estimates to allocate resources, drive new policy decisions or draw conclusions about policy performance – all decisions for which policy makers are (and should be) seeking insights around how much confidence they can place in the small area estimates underpinning their decision-making.

In contrast, the spatial microsimulation approaches that have been developed and applied to date do not provide similar confidence intervals around their central point estimates, in part a reflection of their origins in techniques of geocomputation and simulation rather than statistics and in part a result of methodological challenges around the task. This neglect of

uncertainty around spatial microsimulation small area point estimates is recognised within the literature as the Achilles heel to an otherwise innovative and powerful methodology, undermining its potential and utility for all user groups but particularly for its ability to rigorously inform policy decision-making. Spatial microsimulation scholars are well aware of this weakness and of the pressing need to develop new techniques for the creation of intervals around their central point estimates. Robert Tanton, a key member of the GREGWT spatial microsimulation team in Australia and the broader international spatial microsimulation community, recently recognized this, stating explicitly with colleagues: "This has been *the* biggest difficulty with the modelled small area estimates derived by the ABS [the Australian Bureau of Statistics' GREGWT approach] – there is no estimate of the reliability of the results, for example, standard errors or confidence intervals" (Tanton et al., 2014:80, italics added).

To our knowledge the work of Nagle et al (2014) is the only currently published spatial microsimulation work within the peer-reviewed literature that has attempted to offer central small area point estimates along with accompanying intervals. Hence, from a methodological perspective, there is a significant gap in knowledge around the production of confidence intervals within a spatial microsimulation framework and a need to continue to develop innovative solutions to this key challenge. To do so the paper develops and robustly validates an innovative hybrid statistical-spatial microsimulation approach to the derivation of intervals around IPF small area point estimates.

We demonstrate the proposed method using the IPF technique but the approach can be applied equally to the GREGWT method as both involve, albeit in different ways, the reweighting of national survey data to local small area benchmark totals in what is often described as a deterministic method (i.e. no randomness is involved and the same results are achieved with each run). The proposed approach is not suitable for the conceptually rather different combinatorial optimisation method as that technique involves the use of random number generation within the selection and reselection of survey cases such that the same results are not achieved with each run.

To demonstrate the approach, the paper focuses substantively on the small area estimation of poor health across Wales using survey data from the National Survey for Wales 2013-14 and small area covariate data from the England and Wales Census 2011, contributing to research on the utility of SAE as a census data replacement. The next section describes the IPF approach in greater detail, presents the small area central point estimates and validates these against the Census 2011 data on poor health. This is followed by a discussion of the approach to estimating intervals around these point estimates and consideration of the quality of the resulting intervals. A final section discusses the implications and next steps for the spatial microsimulation community.

3. Small area estimation through spatial microsimulation: iterative proportional fitting in action

Excellent detailed overviews of the IPF approach to spatial microsimulation exist elsewhere (Simpson and Tranmer, 2005; Ballas et al., 2005; Anderson, 2007; Whitworth et al., 2013)

and are only summarised here. The first task within an IPF approach is to identify a survey dataset containing the target outcome of interest as well as a set of predictively useful explanatory variables that are also available as covariate data at the target small area scale. These small area covariate data are, as here, often sourced from Census data, although covariate data may also be available from administrative, commercial or other sources. As noted above, in this paper we focus as our case study on the small area estimation of poor health from the National Survey for Wales 2013-14. Although it would be more usual to focus on the estimation of an outcome not available at small scale, the choice of poor health within a methodologically oriented paper enables us to later conduct rigorous external validation of the IPF estimates and their intervals at the target small area scale using the known poor health data from the Census 2011. Poor health is coded as a binary outcome where those self-reporting in the survey as being in poor health (just under 10% of the cases) are coded one on the outcome and those self-reporting as in good or fair health are coded zero.

A key task is to narrow down the list of potential explanatory factors affecting the poor health outcome to the most parsimonious set of predictively useful factors. Currently researchers take a range of approaches to this task. An initial innovation that we suggest is the formalisation of this task through the use of multilevel multiple regression models of the base survey data to guide decision-making around the optimal set of constraints to use in the IPF based on a balance of predictive power and model parsimony and constrained by small area covariate data availability. In contrast to the work of Anderson (2007) who uses stepwise models focused mechanistically on p-values for this task, we advocate theoretically and empirically guided researcher development of these models.

In our Welsh case study, Table 1 shows the full specification and results from the final individual-level multilevel binary logistic regression model where survey individuals (level 1) are nested inside MSOA small areas that are the target scale for the IPF (level 2), with an average of 33 survey cases in each MSOA. The underlying model specification is as follows:

$$logit(\hat{P}_{ij}) = b_0 + b_{1j} X_{1ij} + \cdots + b_n X_{nij} + u_j, \quad where\ u_j \sim N(0, \sigma_u^2) \qquad (1)$$

The final model offers a reasonably solid foundation for the IPF with a McFadden's pseudo-R2 statistic of 40%, in line with previous occasional studies that have used and presented a comparable statistical approach to constraint selection (Anderson, 2007). These constraint variables are prepared in the base individual level survey file as a set of binary indicator variables and for the small areas as aggregate population counts derived from Census 2011.

Table 1: Multilevel model specification for the estimation of poor health to Welsh MSOAs

Age-Sex	Female 30-49	1.88*	Highest Quals[2]	Level 1	0.85

[2] Level 1 qualifications are equivalent to GCSE grades D-G and NVQ Level 1; Level 2 qualifications are equivalent to GCSE grade A*-C, NVQ Level 2 or Intermediate Apprenticeships; Level 3 qualifications ar

(ref= Female 16-29)	Female 50-64	2.84*	*(ref=no quals)*	Level2	0.85
	Female 65+	1.94*		Level 3	0.87
	Male 16-29	0.56		Level 4+	0.68*
	Male 30-49	1.86*	Health *(ref=no limiting*	Has limiting illness	55.4*
	Male 50-64	2.66*	Region *(ref=North East)*	East	0.85
	Male 64+	1.88*		South-	1.51*
Tenure *(ref=private*	Owned	0.77*		South-East	1.09
	Social Rent	1.27		South-West	1.50*
Employment Status *(ref=unemployed)*	Employed	0.68		North-West	1.19
	Retired	1.67	Constant		0.02*
	Inactive	2.88*			
	Student	0.45			
Observations=13,566 MSOAs (level 2 groups) = 410 Observations per MSOA (level 2 group): min=5; average=33.1; max=114 ** denotes p-value < 0.05*			Explained log-likelihood / Total log-likelihood (McFadden's Pseudo-R2) = 0.40 Residual ICC=0.01 *(Empty ICC=0.04)* Variance of the residual level 2 error=0.0394 *(in empty model=0.1228)*		

This multilevel specification requires the target small area geocodes in the survey file. Although not universally available such small area geocodes are obtainable increasingly on a range of key survey data in the UK context, even if their release often requires the signing of additional data disclosure agreements or secure access. In the case of these survey data, small scale Lower Layer Super Output Area (LSOA) geocodes were included in the survey data and the small area estimation then worked to the slightly larger Middle Layer Super Output Area (MSOA) geography into which LSOAs nest and to which geocodes were aggregated. There are not sufficient survey sample sizes within these geocoded base surveys to estimate directly to the target small area scale, indeed there are areas with no survey respondents. Hence the continued need for SAE techniques despite knowing the small area geocodes of the survey cases.

It is worth clarifying briefly at this point the advantages of advantages of an IPF spatial microsimulation approach to the SAE when it is conceptually possible for the analyst to also progress from here with a regression-based approach. Firstly, the spatial microsimulation approach enables the creation of a synthetic population micro-dataset comprised of multi-way cross-tabulated individuals. This dataset can be used for further analyses such as distributional estimates of the target outcome for small areas or the small area impact of 'what if' policy scenarios or it can be usefully linked to other datasets or simulation models (Vidyattama et al., 2015). In contrast, regression-based approaches struggle to incorporate this individual-level granularity because of the limited availability of individual level census data for reworking models to produce estimates for all areas. As Twigg et al. (2000) note, in one of the methodologies to address this constraint, the required census data are seldom available beyond three-way cross-tabulations. In this context, potential gains from spatial

equivalent to A-levels, NVQ Level 3 or Advanced Apprenticeships; Level 4 qualifications and above include Degrees, Postgraduate Qualifications and Higher Apprenticeships.

microsimulation approaches highlight the importance of delivering confidence intervals around estimates. Secondly, Tranmer et al (2005) argue that spatial microsimulation allows for the complex multilevel structure, and interactions, of individuals, households and localities to be incorporated into SAE analyses.

Once the data are prepared the IPF can be implemented across each target small area in turn. The IPF begins with the initial survey weights and its task is to move across the pre-identified constraint variables in turn and each time to fractionally reweight the survey cases on that constraint according to the extent to which the aggregated weighted values on the survey cases on that constraint variable either over-represent or under-represent that characteristic in the small area. The explanatory factors identified here become the set of constraints to go into the IPF. Formally, the weights on each survey case are reweighted on each constraint according to the following formula,

$$w_{ijk} = w_{ijk-1} * (C_{jk} / S_{jk}) \quad (2)$$

where C_{jk} is the small area aggregate count of constraint *k* in small area *j* (taken typically from Census tables), S_{jk} is the survey weighted sum of constraint *k* in small area *j* based on the most recent survey reweight, w_{ijk-1} is the weight relating to survey case *i* in small area *j* from the previous constraint reweighting, and w_{ijk} is the resulting new weight for survey case *i* in small area *j* from the current reweighting on constraint *k*.

The reweighting technique can be demonstrated with the help of a worked example. Let us assume that the weighted survey total shows 2500 individuals with limiting illness but the target small area contains only 200 individuals with limiting illness. The weights for survey individuals with health conditions will be refined downwards based on the ratio between the two (200/2500=0.08). Hence, the extent to which this deflation of the weights occurs for these survey respondents varies according to their differing needs in terms of replicating the target small area population profile for each group on this constraint variable.

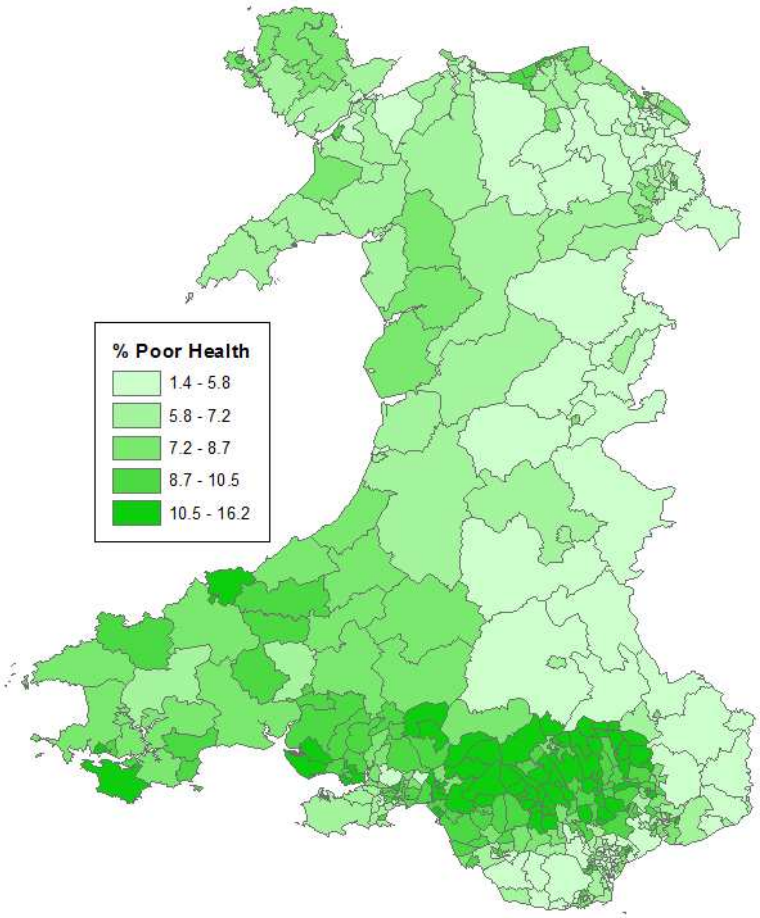
The new, deflated weights then become the starting point for the further reweighting on the next constraint (e.g. economic activity), and so on across each constraint. By doing so the weights are gradually refined as the IPF moves across each of the constraint variables in turn, bringing the weighted aggregated profile of the survey dataset gradually closer both to the size and multi-dimensional profile of the small area population. The most powerful predictive factor (limiting illness) is used as the last constraint in order to maximise its fit. In our approach the IPF sequentially loops around the set of constraints ten times in order to make increasingly fine adjustments to the weights such that they stabilise.

The final calculated weight variable shows the specific weighting that each survey case takes for that small area in order for the survey cases taken as a whole to optimally fit the multi-dimensional profile of each small area. It is then a trivial task to create an estimate of the target outcome variable(s) for each small area by taking a weighted total of the outcome variable across the survey cases. Typically this weighted small area estimate is a point estimate such as a weighted mean or median but distributional estimates of the target outcome variable can also easily be calculated.

A final necessary step in the process is to validate the small area estimates both externally in terms of the face validity of the estimates and internally in terms of goodness of fit on the constraints. Understandably, external validation is often challenging given that comparable small area data often do not exist given the need for SAE in the first instance. In this paper's example, a key reason for estimating poor health as the outcome variable is that this can be validated at the target small area level given that the variable is collected in the UK Census. Across Wales's 410 MSOAs the Pearson's correlation coefficient between the Census percentage of adults in poor health and the equivalent IPF estimates shown in Figure 1 is extremely strong at 0.93. This does not necessarily show that they lie along a 45-degree line starting from the origin as one would ideally like, however, and simple bivariate linear regression can be used to explore this (Scarborough et al, 2009; Taylor et al., 2016): if the estimation has produced perfect results then the intercept of this regression model should be estimated as zero and the single coefficient should be estimated as one. For these poor health estimates the intercept in this model is 0.405 (suggesting that are the IPF estimates tend on average to be 0.4 percentage points higher than the Census percentages), the coefficient is estimated as 1.063 (suggesting only a slight deviation in slope from the ideal 45-degree line) and the adjusted R-square is 0.85. The internal validation is highly effective on the fitted constraints and acceptable on non-fitted constraints using standard fit statistics (Smith et al., 2011). Most fitted constraints give mean standardized errors (MSEs) of zero and virtually all produce MSEs of 0.3 or below. All target small areas have IPF reweighted counts within 20% of the actual Census counts. Five non-fitted constraints were also assessed: being higher, medium or manual socio-economic status; having access to a car; and having dependent children. The IPF performed relatively well here too with MSEs of 10.5, 13.6, 13.1, 6.6 and 10.6 respectively. Taken together these external and internal validation statistics provide strong evidence at the detailed target small area scale for the effectiveness of this small area estimation.

Figure 1 shows the resulting IPF small area point estimates of the percentage of adults estimated to be in poor health across the target Middle Layer Super Output Area (MSOA) scale across Wales, areas with an average population size of 7,860 residents.

Figure 1: Small area estimates of poor health across Welsh MSOAs



4. Getting confident in spatial microsimulation: a new approach to estimating credible intervals

Although analysts using IPF rightly highlight the importance of the validation of point estimates, the process of IPF (and indeed, all forms of spatial microsimulation) currently ends with point estimates. This is deeply problematic for the wide range of users of the resulting small area estimates – policy makers, commercial organisations, charities, academics, general public, and so on – who require information not just about the central point estimates but also crucially about the likely range of values in which the 'true' (but unknown) population value can be expected to fall. This is key additional information to enable users to evaluate how much credence they wish to place on the estimates and what types of business, policy or financial (e.g. resource allocations) they are, and perhaps are not, prepared to make on their basis.

Spatial microsimulation researchers are well aware of this critical weakness and have been explicit in describing an urgent need to make progress in the creation of intervals around their central point estimates (Tanton et al., 2014:80). Initial attempts made using Bayesian approaches offer potential (Rahman et al., 2010) but are not fully developed or tested and face acknowledged challenges in obtaining suitable prior distributions for interested events. Nagle et al's (2014) work on dasymetric modelling, entropy and downscaling offers an alternative approach and one that is to our knowledge the only currently published

methodological approach in this context. Intriguingly, and helpfully at this stage of methodological development around this key gap in the literature, it is distinct from our own proposal for an innovative hybrid statistical-spatial microsimulation approach for the calculation of credible intervals around spatial microsimulation point estimates. We hope that our proposal and that of Nagle et al will further stimulate collective debate and activity across the microsimulation research community.

To this end, the underlying regression model presented above in Table 1 can be further harnessed to open the pathway towards the derivation of confidence intervals around the point estimates following an approach utilised in the statistical SAE literature drawing on the residual between-area error term (Heady et al., 2003; Bajekal et al., 2004; Pickering et al., 2004). In single-level regression specifications the total variance in the outcome variable is assessed at a single level and R-square statistics are customarily used to describe model power in terms of the share of that total variance that can be accounted for by the explanatory factors in the model. In a multilevel regression specification, by contrast, the total variance in the outcome is partitioned across the (two or more) levels of the hierarchy, denoted in a two-level multilevel specification via the intra-class correlation coefficient (ICC) and variance terms at each level in the model. The incorporation of explanatory variables into the multilevel regression model enables the total variance in the outcome to be accounted for separately across the various levels of the model and therefore delivers estimates of residual error at each level of the multilevel structure, as well as of the variance around those residual error terms. In a small area estimation context it is confidence in the precision of the area level point estimates, and a desire to discriminate confidently between point estimates across different small areas, that is of interest. As such, within the two-level multilevel model presented above in Table 1 it is the estimated variance on the residual between-area (i.e. level two) error at the target small area scale that offers the key information for the construction of credible intervals. The greater our ability to account for the between-area variation in this multilevel model and the lesser the extent of the remaining uncertainty at the area level then the tighter can be, and should be, the intervals around the central small area point estimates.

As such, the understanding of 'optimality' is opened out to two separate dimensions against which the underlying modelling endeavours to deliver. A first and more standard understanding of optimality relates to the predictive power of the model and resultant expectation of accuracy in the small area point estimates with a parsimonious set of constraint variables. In terms of the width of the credible intervals, however, a second dimension of optimality relates to the ability within the multilevel specification to explain the between-area variance across the data and, as a result, to narrow the width of the resulting intervals. As such, it is *in principle* possible for a set of modelled explanatory factors to produce underlying models that are sub-optimal in terms of the first dimension of predictive power but that are nevertheless optimal in terms of the second dimension of minimization of the residual between-area variance, and vice versa.

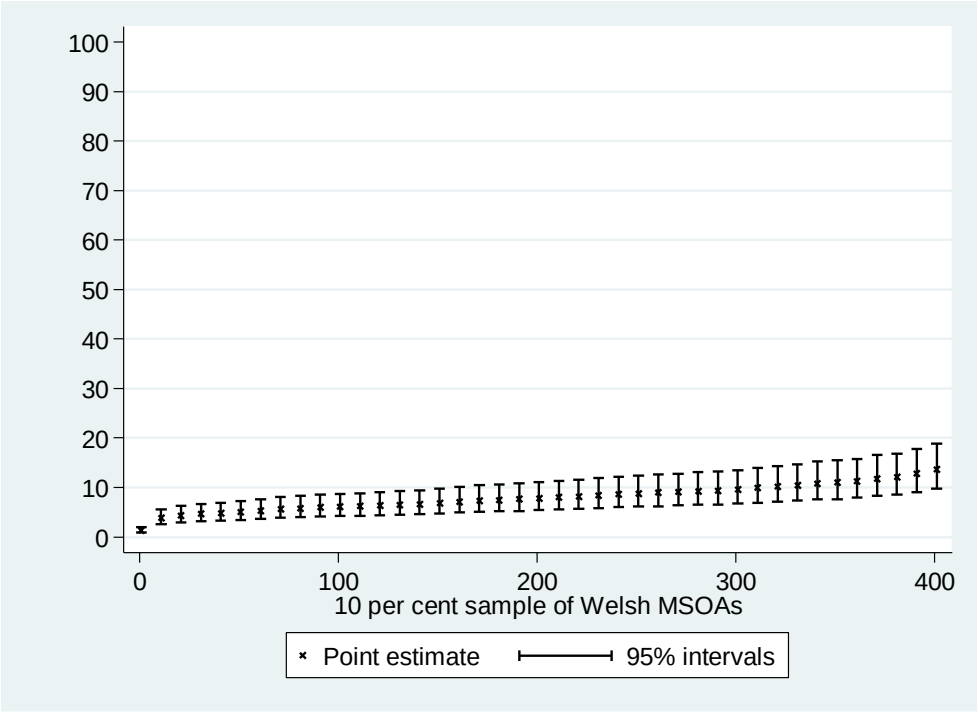
Applying this to our worked example of the small area estimation of poor health across Welsh MSOAs, the estimated standard deviation of the residual between-area variation in

the underlying multilevel binary logit model is shown to the bottom-right of Table 1 above. The shape of this residual between-area error term is now known: its standard deviation is estimated; its mean is assumed to be zero; and its normality is ordinarily assumed, and in this example has also been verified empirically. As such, a distribution of the residual between-area error can be drawn and utilised in order to give a sense of the likely uncertainty around those IPF point estimates.

The process of utilising this information in order to compute the intervals is as follows. For each target MSOA the IPF reweighting delivers a central small area point estimate of the percentage of adults in poor health for each Welsh MSOA. This small area estimate, however, fails to take into account the uncertainty around it. Therefore, for each small area 10,000 separate values are then drawn randomly from the known distribution of the residual between-area error term as described above with mean of zero, standard deviation as estimated by the multilevel model containing the constraints used in the IPF, and normally distributed. The central point estimate and the 10,000 separate between-area error terms are expressed as log odds. Each randomly drawn between-area error term is added separately to the central point estimate for that small area to produce 10,000 plausible small area estimates, each combining the small central point estimate with a slightly different value on the between-area error term that is added. These estimates, now taking into account uncertainty, can then be converted from predicted log odds into predicted probabilities and the 95% credible intervals can be picked off from the 2.5^{th} percentile and the 97.5^{th} percentile of the distribution of these 10,000 separate plausible estimates. Figure 2 provides a visual summary of the resulting credible intervals around the IPF point estimates presented above across a ten per cent sample of Wales' 410 MSOA areas. In keeping with the nature of their calculation we term our results 'credible intervals', a terminology that is standard in the statistical literature.

This is the open access pre-publication version of the article published in Computers, Environment and Urban Systems (2016). Please cite the CEUS published version.

Figure 2: Credible intervals around a sample of Welsh MSOA IPF estimates of poor health



A key reason for choosing to estimate poor health for the purposes of this methodological work is its ability to be robustly externally validated at the target small area scale against known Census data. Given that the actual percentage of adults in poor health across MSOAs is known from Census 2011 data it is possible to assess what percentage of those values lie within the IPF intervals. The ability to capture these population values is in a sense the core function of the intervals and hence offers a useful indicator of their performance.

Typically one would focus on the performance of the standard 95% intervals (+/- 1.96 standard deviations around the mean) but it is possible to be more comprehensive in the assessment of the intervals by instead considering the performance of the estimated credible intervals across their entire full distribution. Table 2 offers this more detailed analysis. Specifically, it is possible to take a variety of differently specified levels of standard deviations around the mean and to set out the percentage of cases that one would expect to fall within – and, hence conversely, beyond – these bounds. This expected performance is shown in column two of Table 2 in relation to the variety of standard deviation levels shown in column one. For example, one would *expect* 68.3% of Welsh MSOAs to have 'true' Census 2011 values for the percentage of residents in poor health within one standard deviation, and 95.5% within two standard deviations, of the mean on the estimated distribution of the credible intervals. Column three shows the *actual* percentage of 'true' Census 2011 values that fall within these various bounds based on a comparison of those known Census values against the estimated distribution of the credible intervals derived. The final column shows the ratio between these two (i.e. actual percentage/expected percentage) such that a value of one would mean that the performance of the estimated credible intervals was perfectly in line with expectations.

Table 2 shows that the proposed methodology to derive the credible intervals performs extremely well and matches closely what would be expected across the full range of the distributions of the resulting intervals. Indeed, the estimated intervals here perform slightly better than would be expected at thresholds closer to the mean and by +/- 1.5 standard deviations and beyond their performance is near identical to what would be expected. This is strong evidence of their functionality.

Table 2: Actual and expected performance of the credible intervals

Standard Deviations	Expected % Census values within bounds	Actual % Census values within bounds	Ratio
+/- 0.5 σ	38.3	41.2	1.10
+/- 1.0 σ	68.3	71.7	1.05
+/- 1.5 σ	86.6	87.3	1.01
+/- 2.0 σ	95.5	96.3	1.01
+/- 2.5 σ	98.8	98.5	1.00
+/- 3.0 σ	99.7	99.3	1.00

5. Conclusion

Despite the existence of national Census data in most national contexts and the growing interest in, and availability of, 'new' and 'Big' data sources, widespread gaps continue to exist in the spatial resolution at which key variables of interest exist. Within this context SAE techniques of various forms can be utilised to fill some of those informational black holes, squeezing additional value from existing survey data investments and offering new spatially detailed data insights where they could not otherwise be obtained. Such SAE techniques currently rely on and can supplement Census data and, in the UK context at least, take on an additional future importance given the on-going push away from the traditional Census in this context.

The present paper has focused on spatial microsimulation approaches to small area estimation and the continued inability of those approaches to deliver robust intervals around their small area point estimates. The continued absence of such intervals from spatial microsimulation approaches to SAE seriously undermines the utility of these otherwise powerful methodologies for the various user communities seeking to make use of the additional spatial detailed understanding. This limitation is particularly acute for policy makers who are often the key group requesting the use of small area estimation techniques to deliver for spatial detailed information to underpin their work but whom inevitably also wish to reflect on the likely precision of the point estimates before making decisions around policy interventions or resource allocations.

The paper has presented an innovative hybrid statistical-spatial microsimulation approach to the construction of credible intervals around small area point estimates from spatial microsimulation SAE techniques, based on the IPF estimation of adults in poor health across Welsh MSOAs. The proposed method can be applied either to IPF or to GREGWT spatial

microsimulation approaches. The approach involves the incorporation of a multilevel regression model in the base survey file in order to identify the optimal constraints for the IPF reweighting in a more rigorous and systematic way than is typically the case in the literature at present, with survey individuals nested inside the target small area scale (here MSOAs). Drawing on work in the statistical small area estimation community, given that the chief concern is a desire to discriminate confidently between point estimates across different small areas, then it is the residual variance on the between-area error term that is of key importance within this estimated multilevel model for the derivation of the intervals. With the key characteristics of this residual between-area error distribution known – mean, variance, shape – then it is possible to draw randomly a series of (in our example 10,000) additional error terms with which to add to the IPF derived central point estimates in order to, in effect, perturb the small area estimates according to the estimated extent of their likely precision. The 95% credible intervals can then be picked off from the 2.5^{th} percentile and the 97.5^{th} percentile of the resulting distribution of outcome estimates.

By selecting poor health as the outcome variable, the analyses are able to validate the point estimates and their intervals using the collected Census data of this same poor health variable and at the same target small area scale. Our proposed approach performs extremely well in this worked example. The central IPF point estimates of adults in poor health correlate highly with the Census percentages across Welsh MSOAs (r=0.93) and in linear models produce a near-perfect slope estimate (b=1.063), though with a slightly high intercept estimate (a=0.405). The internal validation is highly effective on the fitted constraints and acceptable on non-fitted constraints.

In terms of the paper's key focus on the derivation of the intervals, the validation is again able to be conducted robustly at the target MSOA scale against the known Census 2011 data. At the standard 95% threshold 96.3% of Wales' 410 MSOAs show 'true' Census values for the percentage of residents with poor health that are within the 95% intervals estimated using our proposed approach. The analyses also examine the performance of the estimates across a series of standard deviation thresholds across the full range of the estimated intervals. At all points throughout this distribution the credible intervals perform extremely well against what would be expected at each level. Our proposed innovative methodology to derive credible intervals in spatial microsimulation SAE approaches therefore appears highly effective and represents a significant step forwards in resolving this key weakness of these otherwise powerful methodological approaches. We call on the broader spatial microsimulation community to pick up this and related work so that we can collectively continue to make progress in the robust estimation of uncertainty around our small area point estimates until such time as they are produced as a matter of course. Only then in our view will spatial microsimulation approaches really have the statistical robustness desired and expected for a small area estimation methodology that can be used by policy makers, business users, third sector groups and the general public in understanding and seeking to improve social and economic outcomes at fine spatial scales.

References

Anderson, B (2007) Creating *small area income estimates for England: spatial microsimulation modelling*, a report to the Department of Communities and Local Government. London: Department of Communities and Local Government.

Bajekal, M., Scholes, S., Pickering, K. and Purdon, S. (2004) *Synthetic estimation of healthy lifestyle indicators: Stage one report*. London: National Centre for Social Research.

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(1), pp13-34.

Ballas, D., Clarke, G. and Wiemers, E. (2006) 'Spatial microsimulation for rural policy analysis in Ireland: the implications of Cap reforms for the national spatial strategy', *Journal of Rural Studies*, pp367-378.

Birkin, M. and Clarke, G. (2011) 'Spatial microsimulation models: a review and glimpse into the future', *Population Dynamic and Projection Methods: Understanding Population Trends and Processes Volume 4*. London: Springer.

Bishop, Y., Fienberg, S and Holland, P. (1975) *Discrete multivariate analysis: Theory and Practice*. MIT Press.

Chambers, R and Tzavidis, N (2006) 'M-quantile models for small area estimation', *Biometrika,* 93, pp255-268.

Chatterjee, S., Lahiri, P. and Li, H. (2008) 'On small area prediction interval problems', *The Annals of Statistics*, 36, pp1221-1245.

Ghosh, M. and Rao, J. (1994) 'Small area estimation: an appraisal', *Statistical Science*, 9(1), pp55-76.

Heady, P., Clarke, P., Brown, G., Ellis, K., Heasman, D., Hennell, S., Longhurst, J. and Mitchell, B. (2003) *Model based area estimation series No. 2: Small area estimation project report*. London: Office for National Statistics.

Hermes, K and Poulson, M (2012) 'Current methods to generate synthetic spatial microdata using reweighting and future directions', *Computers, Environment and Urban Systems,* 36, 281-290.

Marshall, A (2012) *Small area estimation using ESDS government surveys – An introductory guide* Economic and Social Data Service.

Nagle, N., Buttenfield, B., Leyk, S and Spielman, S. (2014) 'Dasymetric Modeling and Uncertainty', *Annals of the Association of American Geographers*, 104(1), pp80-95.

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

Pickering, K., Scholes, S and Bajekal, M (2004) *Synthetic estimation of healthy lifestyles indicators: Stage 3 report.* London: NatCen.

Rahman, A (2008) *A review of small area estimation problems and methodological developments.* University of Canberra: NATSEM Discussion Paper Issue 66.

Rahman, A., Harding, A., Tanton, R. and Liu, S. (2010) 'Methodological issues in spatial microsimulation modelling for small area estimation', *International Journal of Microsimulation*, 3(2), pp3-22.

Rao, J (2003) *Small Area Estimation.* New York: Wiley

Rao, J (2005) 'Inferential issues in small area estimation: some new developments', *Statistics in Transition*, 7(3), pp513-526.

Rees, P, Martin, D and Williamson, P (2002) (ed), *The Census Data System*. Chichester: Wiley

Scarborough, P., Allender, S., Rayner, M and Goldacre, M. (2009) 'Validation of model-based estimates (synthetic estimates) of the prevalence of risk factors for coronary heart disease for wards in England', *Health & Place*, 15(2), pp596-605.

Simpson, L. and Tranmer, M. (2005) 'Combining sample and census data in small area estimates: iterative proportional fitting with standard software', *The Professional Geographer*, 57(2), pp222-234.

Smith, D., Pearce, J. and Harland, K. (2011) 'Can a deterministic spatial microsimulation model produce reliable small-area estimates of health behaviours? An example of smoking prevalence in New Zealand', *Health & Place*, 17(2), pp618-624

Tanton, R and Edwards, K (2013) (eds) *Spatial microsimulation: a reference guide for users*. London: Springer.

Tanton, R., Vidyattama, Y., Nepal, B and Mcnamara, J (2011) 'Small area estimation using a reweighting algorithm', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(4), pp931-951.

Taylor, J., Moon, G. and Twigg, L (2016) 'Using geocoded survey data to improve the accuracy of multilevel small area synthetic estimates', *Social Science Research,* 56, pp108-116

Tanton, R., Williamson, P. and Harding, A. (2014) 'Comparing two methods of reweighting a survey file to small area data', *International Journal of Microsimulation*, 7(1), pp76-99

Tranmer, M., Pickles, A., Fieldhouse, E., Elliot, M., Dale, A., Brown, M., Martin D, Steel D & Gardiner, C. (2005) 'The case for small area microdata', *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 168(1), pp29-49.

Tzavidis, N., Marchetti, S and Chambers, R (2010) 'Robust estimation of small area means and quantiles', *Australian and New Zealand Journal of Statistics*, 52(2), pp167-186.

Vidyattama, Y., Tanton, R. and Biddle, N. (2015) 'Estimating small-area Indigenous cultural participation from synthetic survey data', *Environment and Planning A*, 47(5), pp1211-1228.

Voas, D and Williamson, P (2000) 'An evaluation of the Combinatorial Optimisation approach to the creation of synthetic microdata', *International Journal of Population Geography*, 6, pp349-366.

Whitworth, A (2013) (ed.) *Evaluations and improvements in small area estimation methodologies.* Economic and Social Research Council: National Centre for Research Methods methodological review paper.