Computers, Environment and Urban Systems 63 (2017) 50–57

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

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

A. Whitworth [a,*], E. Carter [a], D. Ballas [a], G. Moon [b]

[a] Department of Geography, University of Sheffield, UK
[b] Geography and Environment, University of Southampton, UK

ARTICLE INFO

Article history:
Received 9 September 2015
Received in revised form 20 May 2016
Accepted 15 June 2016
Available online 21 June 2016

Keywords:
Small area estimation
Spatial microsimulation
Iterative proportional fitting
Credible intervals
Confidence intervals
Variance estimation

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.

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 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 (Chatterjee, Lahiri, & Li, 2008; Rao, 2005) and *the* key (and significant) weakness of spatial microsimulation approaches (Nagle, Buttenfield, Leyk, & Spielman, 2014; Tanton, Williamson, & Harding, 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.

* Corresponding author at: Department of Geography, University of Sheffield, South Yorks S10 2TN, UK.
E-mail address: adam.whitworth@sheffield.ac.uk (A. Whitworth).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.06.004
0198-9715/

2. Methodological approaches to small area estimation

As summarised elsewhere (Bishop, Fienberg, & Holland, 1975; Ghosh & Rao, 1994; Marshall, 2012; Rahman, 2008; Rao, 2003; 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 (Chambers & Tzavidis, 2006; Ghosh & Rao, 1994; Pfeffermann, 2013; Rao, 2003; Tzavidis, Marchetti, & Chambers, 2010). 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 (Anderson, 2007; Ballas, Clarke, & Wiemers, 2006; Birkin & Clarke, 2011; Hermes & Poulson, 2012; Rahman, Harding, Tanton, & Liu, 2010; Tanton & Edwards, 2013; Tanton, Vidyattama, Nepal, & Mcnamara, 2011; Voas & Williamson, 2000). 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 recognised 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.