Social Science Research 56 (2016) 108–116

Contents lists available at ScienceDirect

Social Science Research

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

Using geocoded survey data to improve the accuracy of multilevel small area synthetic estimates

Joanna Taylor [a], Graham Moon [a, *], Liz Twigg [b]

[a] Geography and Environment, University of Southampton, University Road, Southampton, SO17 1BJ, UK
[b] Department of Geography, University of Portsmouth, Buckingham Building, Lion Terrace, Portsmouth, PO1 3HE, UK

ARTICLE INFO

Article history:
Received 24 November 2014
Received in revised form 21 December 2015
Accepted 31 December 2015
Available online 8 January 2016

Keywords:
Multilevel
Synthetic estimation
UK census
Geocodes
spatial identifiers
Limiting long term illness

ABSTRACT

This paper examines the secondary data requirements for multilevel small area synthetic estimation (ML-SASE). This research method uses secondary survey data sets as source data for statistical models. The parameters of these models are used to generate data for small areas. The paper assesses the impact of knowing the geographical location of survey respondents on the accuracy of estimates, moving beyond debating the generic merits of geocoded social survey datasets to examine quantitatively the hypothesis that knowing the approximate location of respondents can improve the accuracy of the resultant estimates. Four sets of synthetic estimates are generated to predict expected levels of limiting long term illnesses using different levels of knowledge about respondent location. The estimates were compared to comprehensive census data on limiting long term illness (LLTI). Estimates based on fully geocoded data were more accurate than estimates based on data that did not include geocodes.

1. Introduction

Statistical approaches to small area synthetic estimation have received significant attention in recent years due to growing demands for consistent, robust and reliable small area data (Scholes et al., 2008; Whitworth, 2012). These demands are not always addressed by national census data, and local surveys do not offer consistent national data. As a result, there remain gaps in the provision of small-area information that are addressed using small area estimation methodologies. These methodologies are a topic of intensely active research (eg. Marchetti, Tzavidis, & Pratesi, 2012; Molina and Rao, 2010; Pfeffermann, 2013).

Area-specific direct estimation using in-area sample measures to draw inferences about population characteristics is rarely possible at the neighbourhood level. This is because national surveys do not normally sample in all localities leading to out-of-sample areas with no respondents on which to base direct estimates. Furthermore in those neighbourhoods that are sampled, sample sizes are seldom large enough to produce reliable estimates (Heady et al., 2003). These difficulties make the case for indirect or synthetic estimates (Chandra, Salvati, Chambers, & Tzavidis, 2012). The basic process behind synthetic estimation can be summarised as "*modelling nationally but predicting locally*" whereby a statistical model is created to predict the expected probability of a 'target variable' using a survey dataset with relevant independent covariate information. Local data are then applied to the coefficients from the national model to generate local small area estimates.

* Corresponding author.
E-mail addresses: j.l.taylor@soton.ac.uk (J. Taylor), g.moon@soton.ac.uk (G. Moon), liz.twigg@port.ac.uk (L. Twigg).

http://dx.doi.org/10.1016/j.ssresearch.2015.12.006

Twigg, Moon and Jones (2000) developed a multilevel modelling approach to (small area) synthetic estimation (ML-SASE) and illustrated their approach through the calculation of electoral ward level estimates of the prevalence of adult smoking and unhealthy alcohol consumption in England. Their approach used data from the Health Survey for England to build multilevel models of smoking and alcohol consumption with independent variables, chosen for their epidemiological relevance and co-presence in both the survey and the UK census. These independent variables were either at the individual level (eg age, sex) or at the area level (eg local deprivation).

Prior to the development of ML-SASE, synthetic estimates were commonly based on statistical models with either solely individual *or* solely area level covariates, whereas the multilevel synthetic estimation methodology incorporated both. The National Centre for Social Research was commissioned by the UK Government's Department of Health to undertake a technical review and evaluate the methodologies for generating small area synthetic estimates of healthy lifestyle behaviours in England. It reported that "*conceptually and methodologically, the analysis by Twigg et al. (2000) represents an innovative advance over the simpler methods … for it accommodates both individual and area level effects*" (Bajekal, Scholes, Pickering, & Purdon, 2004, p. 12). Conceptually including both individual and area level variables in a predictive multilevel modelling framework can avoid both the ecological fallacy (Robinson, 1950) and the individualistic fallacy (Alker, 1969), leading Subramanian et al. (2009, p. 355) to conclude that "*multilevel thinking … is thus a necessity, not an option*".

The importance of this theoretical imperative can be illustrated through the example of predicting the propensity to smoke. A multitude of previous studies have shown that those individuals with a low socio-economic status are more likely to smoke. However, there is also an additional, independent association between the risk of an individual being a smoker and the additional risks that accrue if they live in a neighbourhood with high levels of low socio-economic status individuals who are all more likely to be smokers and hence, arguably, generate a local culture of smoking. Other individual associations with smoking may equally be modified by area level influences. Predicted prevalences for small areas thus need to take into account both individual and area level factors (Duncan, Jones, & Moon, 1999).

The widespread availability of survey data through the provision of data archives has rendered the task of sourcing survey data for synthetic estimation purposes superficially straightforward. However, incorporating both individual and area effects within a ML-SASE framework brings data challenges. In this paper we focus on the importance of respondent spatial identifiers, sometimes referred to as geocodes, within secondary survey datasets – the prime sources of data used for small area synthetic estimation. Such spatial identifiers tell us approximately where each respondent in the survey lives, for example, in England and Wales this may be a code for an electoral ward (a small area local government geography) or a Super Output Area (a small area used in the reporting of census results and other official statistics[1]). Usually, geocodes do not tell us exactly where the respondent lives. The release of household addresses, geographical coordinates or full postcodes is limited in order to ensure respondents' anonymity.

Our aim is to examine quantitatively the implications of varying levels of geocoding for the use of area level data in ML-SASE. We do this by making and comparing different sets of synthetic estimates which, in terms of their methodologies, differ only with respect to the way in which area level data are generated via geocoding. The next section places our aim within the context of the data requirements for the multilevel small area synthetic estimation process and elaborates on the ways in which area level data can be generated. Section 3 outlines the methodology employed to address our research questions and Section 4 compares the resultant sets of synthetic estimates. As well as acknowledging the study's limitations, the concluding section addresses the implications of our results both in terms the choice of the social survey datasets that form the basis for sets of multilevel synthetic estimates and with respect to current and future plans for access to geocoded social surveys.

2. Background – the data requirements for ML-SASE

The first stage to generating multilevel synthetic estimates is to choose a large scale social survey dataset. As Dale (2006) has previously argued, UK researchers are in the fortunate position of having access to many data sets that facilitate the analyses that are needed to determine both individual and area level influences on a vast array of individual outcomes. The UK Data Service currently holds around 6000 data collections covering a wide range of both economic and social data and includes many of the major UK surveys (UK Data Service, 2013). Unfortunately, because of the secondary data requirements for ML-SASE, only a selection of these survey datasets is currently suitable for ML-SASE purposes. For the purposes of this paper, this limitation reflects two broad reasons. These relate to the hierarchical structure required for multilevel models, and to our key focus on the possibilities for including area level explanatory variables. Each merits a brief discussion.

2.1. Hierarchical structures

The hierarchical or multilevel structure of the survey data that is used develop ML-SASE models commonly comprises individuals, nested within small areas, which in turn are sometimes nested within larger geographies such as regions or Local Authorities (Fig. 1). The first hierarchical level is the individual respondent. There is consistent evidence that most of the

[1] Super Output Areas are a small area partitioning of England and Wales used in both the 2011 and 2011 Census covering England and Wales. They come in two sizes, the smaller Lower Layer Super Output Areas (LSOAs) each with a population of between 1000 and 3000 can be amalgamated into larger Middle Layer Super Output Areas (MSOAs) each with a population of between 5000 and 15,000.