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

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.

Keywords

Multilevel, synthetic estimation, UK census, geocodes, spatial identifiers, limiting long term illness

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

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, Pickering, & Deverill, 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 & 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.

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. 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[i]). 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

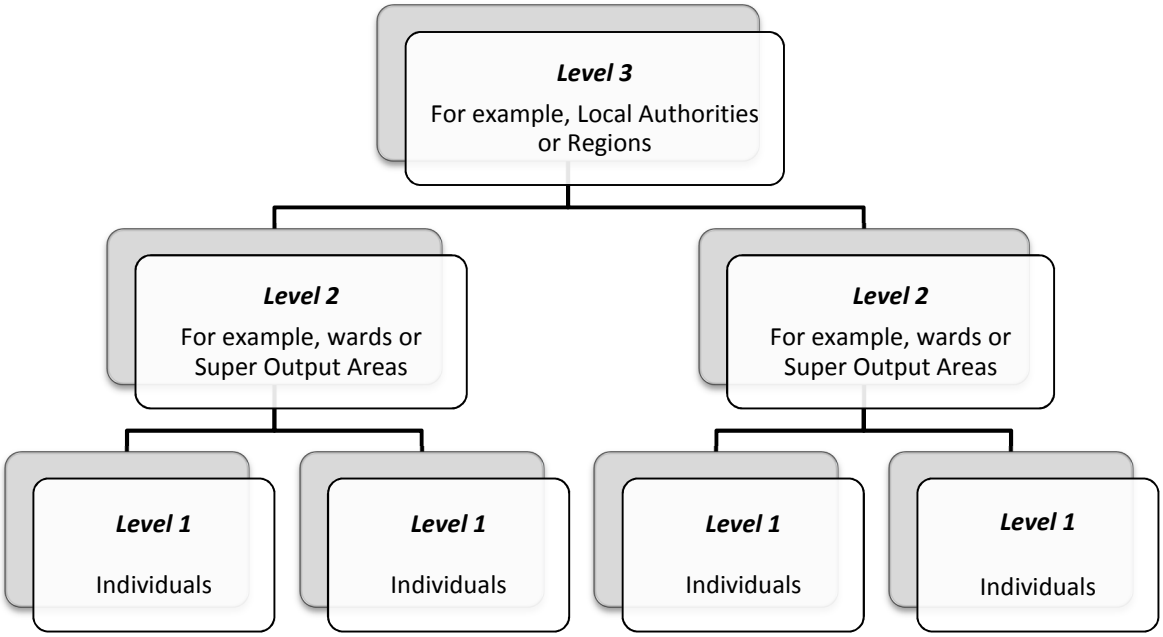
The first stage to generating multilevel synthetic estimates is to choose a large scale social survey dataset. As Dale (2006) has previously argued, UK researchers are in the fortunate position of having access to many data sets that facilitate the analyses that are needed to determine both individual and area level influences on a vast array of individual outcomes. The UK Data Service currently holds around 6,000 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 (Figure 1).

The first hierarchical level is the individual respondent. There is consistent evidence that most of the variance in datasets used for synthetic estimation is at the individual level – examples of the percentage of variance at this level from previous multilevel models used to generate synthetic estimates include smoking (93%), obesity (97%) and fruit and vegetable consumption (90%) (Pickering, Scholes, & Bajekal, 2004) and various fear of crime indicators (91% to 94%) (Whitworth, 2012). This substantiates the need to include an individual level in the ML-SASE process. This case is furthered by the conceptual advantage of incorporating level 1 to provide an equivalence to age-sex standardisation. Moreover, the inclusion of level 1 variables makes it feasible to calculate subgroup estimates (Scholes et al., 2008).

Figure 1 An example of a hierarchical data structure (based on England and Wales geographies)



Level two is the first and lowest of the levels relating to a geographical area; it is conventionally the geography for which the small area synthetic estimates are required or from which they can be built. Surveys commonly do not include a variable indicating exactly where respondents live in order to ensure respondents' anonymity. In certain instances however, the survey dataset may include a variable indicating which respondents reside in the same cluster unit or primary sampling unit used in the sample design of the survey. In this way it may be possible to know which respondents live in some proximity to each other, if

not their actual location. Unfortunately, primary sampling units do not necessarily translate to the geography required for the synthetic estimates; the target spatial unit for estimation needs to be at the same or near equivalent scale to the area level employed in the modelled survey.

A drawback of many approaches to synthetic estimation is that they are design-biased, meaning that, for a particular area, they estimate the underlying *expected* value for any area given the socio-demographic independent variables included in the model rather than *real* value for the small area in question (Heady et al., 2003). To reduce the design bias of the estimates, survey data sets for modelling should ideally possess a third level in their hierarchical multilevel structure; the level 2 small areas should nest within larger geographical areas such as regions. To be useful for estimation purposes there should be respondents in all of these larger (level 3) areas. This allows the fixed effects of variables included in the model to be supplemented by a residual for each level 3 unit, thus reducing the design bias of the synthetic estimates. Twigg, Moon and Walker (2004) used Government Office Region level residuals in this way to improve their estimations of smoking prevalence.

In the light of these requirements, the ideal scenario is to have access to social survey datasets with spatial identifiers attached to each respondent's record. Spatial identifiers such as census geography codes, geographical coordinates, or postal codes enable, either in their own right or via look-up tables, the development of models in which researchers are able to define the geographical hierarchy themselves taking into account the clustering in the sampling strategy, the numbers of respondents in the small areas, and the geography required for the synthetic estimates from their end user(s).

2.2 Area level explanatory variables

In terms of generating area level explanatory variables, researchers have three options. ***Option A*** is to derive area data based solely on the characteristics of survey respondents in the survey who live in each area. For example, if the percentage that lives in private rented tenure across small areas was an important area-level covariate in the model then the proportion of respondents declaring such tenure across each of the small areas in the survey could be used as an area level variable. This is facilitated in situations where the researcher knows which respondents form a sampling cluster within the survey; cluster membership may be disclosed but not the actual geographical location of the cluster. It is also of course possible in cases where respondents' geographical areas of residence are known.

Any area level estimates derived from aggregating individual responses in this way to represent an area as a whole is obviously likely to be imprecise given small sample sizes. For example, the 2010/11 sweep of the Crime Survey for England and Wales contains 3,707 Middle Layer Super Output Areas with an average of 12.6 respondents. If a quarter of this average number of Crime Survey respondents in a Middle Layer Super Output Area said they rented privately, the true area percentage (based on 95% confidence intervals) would be between zero per cent and 53.7 per cent. Taking an alternative example, the 2011 sweep of the Health Survey for England has a variable indicating which respondents belong to the same primary sampling unit but not the geographical location of the units. On average there were 18.9 respondents in each primary sampling unit with a mean age of 42.3. The true area mean age would be between 28.1 and 56.6 years old.[ii] Both examples highlight the problem with extrapolating aggregated individual level respondents' answers to represent the entire area.

Option B is to utilise the limited area level data now routinely attached to many large scale UK social surveys even when actual small areas are not disclosed. For example, the Health Survey for England includes an indicator of rurality (The Countryside Agency, Department for Environment Food and Rural Affairs, Office of the Deputy Prime Minister, Office for National Statistics, & Welsh Assembly Government, 2004) as well as quintiles of the Index of Multiple Deprivation (IMD) (McLennan et al., 2011) attached to each respondent via their undisclosed location. The Crime Survey for England and Wales has, amongst others, the different domains of the IMD and two geodemographic typologies (the Output Area Classification (Vickers & Rees, 2007) and the Classification of Residential Neighbourhoods (CACI, 2009)). The British Social Attitudes survey contains area information on population density for each respondent's area of residence.

These routinely attached data are often in conflicting spatial formats, measured at different spatial scales. Moreover, for ML-SASE purposes, the area information may be for small areas that do not reflect the sampling units in the survey design. For example, the Health Survey for England provides the IMD quintile for the Lower Super Output Area but its sampling design was based on postcode sectors. In addition, although routine attachment of area level data is increasing, the number and range of such variables is driven by survey sponsors' requirements. As a consequence, potentially significant area level variables may not be

appended routinely. An example of this, in the context of estimating neighbourhood health status, is the UK Department for Work and Pensions' neighbourhood data on the levels of claimed benefits. Though previously shown to be associated with morbidity (Henderson, Stansfeld, & Hotopf, 2013; Norman & Bambra, 2007), these data are not routinely attached to the Health Survey for England.

Option C is for researchers to be able, independently, to link appropriate area level variables (referred to in Figure 1 as level 2) to respondents' survey records. In contrast to Option A, the linked data are ecological area variables rather than aggregates of individual variables. There are two ways to approach Option C: the first is to request the survey company originally responsible for the base survey to attach additional area variables – however, this often has cost implications. The alternative is to gain access, possibly via secure settings, to versions of the surveys with small area geocodes in the form of postcodes or census geography codes. The researcher can then append the desired area level data. This option offers great flexibility, in that the researcher can choose which data to link, but carries with it significant risks regarding disclosure as the provision of geocodes in conjunction with survey data may enable the identification of individual survey respondents. For this reason secure access arrangements are generally essential. Methodologically Option C has advantages in that it offers the possibility of attaching Census data which provide a more comprehensive summary of an area, one based on all its residents rather than only those completing a sample survey.

The three options can be summarised using simple formulae in which a target area-level variable (A), for j areas can be created from combination of numerator (N) and denominator (D data at area-level (j) or the individual-level (i), or from appended indicator data (ID) at the area-level (Table 1)

Table 1 A Comparison of Area-level Measurement Options

Option A	**Option B**	**Option C**
Aggregating individual data from survey	Using provided area indicators	Appending external census data
$A_j = \frac{\sum N_i}{\sum D_i}$	$A_j = ID_j$	$A_j = \frac{N_j}{D_j}$

3. Materials and Methods

We seek in this paper to examine quantitatively whether knowing the (geocoded) area of respondents' residences can improve the accuracy of multilevel synthetic estimates. More specifically we test whether:

(1) Having independent information about the local area though knowing the actual area of residence of respondents (Option C) can improve the accuracy of synthetic estimates compared with aggregating respondents' answers to provide survey based estimates of local area characteristics (Option A)?

(2) Having more detailed information about the local area through knowing the area of respondent residence (Option C) can improve the accuracy of synthetic estimates compared with utilising summary area information routinely added to the survey dataset by survey contractors (Option B)?

To address our two research questions we generated four sets of multilevel synthetic estimates for the percentage of the adult population with a limiting long term illness (LLTI) for every Middle Layer Super Output Area in England using the methodology developed by Twigg, Moon and Jones (2000). We measured LLTI using the standard UK social survey question: are your day-to-day activities limited because of a health problem or disability which has lasted or is expected to last for more than 12 months (include problems related to old age). LLTI was chosen as the exemplar for two reasons. First, small area estimates of morbidity offer important insights into local scale variations in health need and demand for health and social care services (ONS, 2010). Second, a question on LLTI was also asked in the 2011 UK Census thus providing an alternative localised data source, surveying (in theory) all residents in those localised areas and offering a 'gold standard' measure of the small area prevalence of LLTI

The survey dataset used to generate the ML-SASE models was the 2010/11 sweep of the Crime Survey for England and Wales. At the time of writing a geocoded version with respondent spatial identifiers was available via a special licence from the UK Data Service. The Crime Survey for England and Wales has the advantages of a relatively large sample size (46,754 with a response rate of 76 per cent) and the fact that the primary sampling units are

based on the census geography of Super Output Areas, with the sampling process being stratified by Police Force Area within a partially clustered design (Fitzpatrick & Grant, 2011). All our multilevel models included age (16 categories) and sex (two categories) at the individual level. We supplemented these covariates with additional area level variables of ecological deprivation reflecting our different geocoding scenarios. Numerous studies have demonstrated the association between ecological deprivation and poor health (see for example Dorling & Thomas, 2004).

To address research question one, ecological deprivation was quantified as the percentage of the people living in the local area who were unemployed. This was either calculated as the number of unemployed crime survey respondents in each small area as a percentage of the total number of respondents in each area (Option A) or as the number of residents recorded as unemployed in the 2011 Census as a percentage of the total resident 2011 Census population. This census-based measure was linked to the survey data via geocoding (Option C). To address research question two, area level deprivation was quantified using the English Index of Deprivation (ID) (McLennan et al., 2011) either using the deciles that are routinely appended to the survey for each area (Option B) or the full raw ID score for the area, which we appended to the survey using geocoding (Option C). Table 2 summarises the area measures in the four comparison models.

Table 2 Summary of the four approaches to modelling area level deprivation

Research question (1) **Area unemployment**		
Option A	***versus***	**Option C**
Percentage of Crime Survey for England and Wales respondents unemployed in each Middle Layer Super Output Area.		Percentage of residents unemployed from the 2011 Census in each Middle Layer Super Output Area.
Research question (2) **Area deprivation**		
Option B	***versus***	**Option C**
Index of Deprivation deciles already attached to each respondent in the survey.		Index of Deprivation score (excluding the health domain).
Middle Layer Super Output Area indicator expressed as decile membership for national ranking of Middle Layer Super Output Areas.		*Middle Layer Super Output Area indicator calculated as the population weighted average of the scores from the Lower Layer Super Output Areas in each Middle Layer Super Output Area.*

The multilevel models were generated using MLwiN v2.30 (Rasbash, Browne, Healy, Cameron, & Charlton, 2014) and were estimated via Markov chain Monte Carlo simulation. The default prior distribution applied by the software package for all the parameters was flat. Information on the calculation of conditional posterior distributions can be found in Browne (2012). The model was run through 500,000 iterations (with a burn in period of 50,000 iterations). The Raftery-Lewis diagnostic (Raftery & Lewis, 1992) and the Effective Sample Size (Kass, Carlin, Gelman, & Neal, 1998) both confirmed that this Markov chain length was sufficiently long.

To assess model quality we examined differences in model fit using the Deviance Information Criterion (DIC) (Spiegelhalter, Best, Carlin, & A. van der Linde, 2002). This can be thought of as a measure of how well the model fits the data, with a lower score suggesting a better model. We also examined variations in the percentage of variance explained by the different models using the standard approach defined by Snijders and Bosker (2012, 306).

We compared the synthetic estimates generated under our different options with census data on LLTI using scatter plots of the synthetic estimates (x axis) against the 2011 Census (y axis). A first diagnostic was to investigate whether the synthetic estimates and the census data were strongly positively associated via Spearman's rank correlation test. This test was chosen to explore the correspondence in the rank ordering of the two measures. A second more stringent test examined a regression line fitted to the scatterplots of the two sets of estimates. Although we would expect there to be a wide scatter (due to confidence intervals around the synthetic estimates), a close correspondence between the synthetic estimates and the gold standard census data should result in a scatter around the line $x = y$, in other words a regression line with a gradient close to one and an intercept around zero (Scarborough, Allender, Rayner, & Goldacre, 2009).

4. Results

Table 3 compares the results of the four multilevel models. In relation to research question one, the model based on the geocoded version of the dataset (Option C) has a DIC statistic 979 points less than the corresponding model based on aggregated respondent information (Option A). Similarly with respect to the second research question the DIC statistic is 68 points lower for the model with full linked IMD data than for the corresponding model based

on the Index of Deprivation deciles already appended to the survey dataset by the research company (Option B). Because a fall of ten or more in the DIC statistic conventionally indicates that the worse model (the one with the highest DIC statistic) has less support (Spiegelhalter, Best, Carlin, & van der Linde. (2002), we can conclude that, based on the DIC statistic, the two Option C models represent a better fit of the data. The two Option C models also explained a higher proportion of the model variance (1.9 percentage points more in relation to research question one and 0.8 percentage points in relation to research question 2).

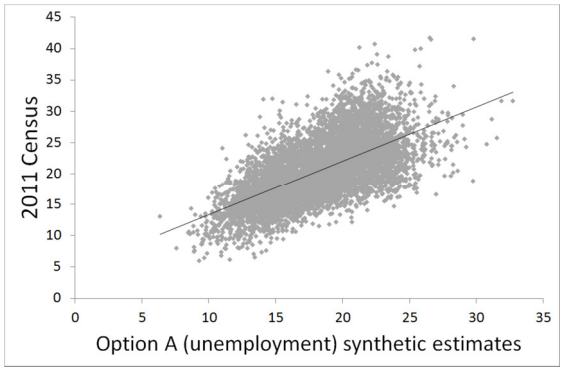
Table 2 Comparing the multilevel models

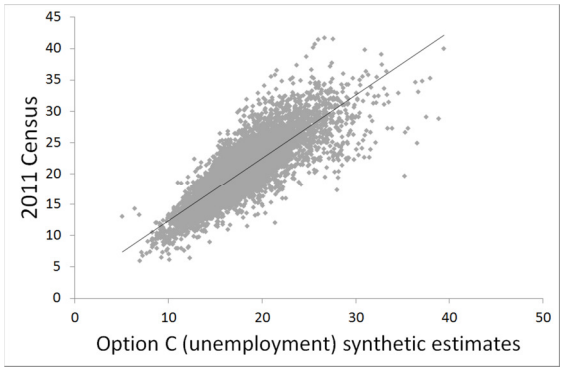
	Level 2 variance (95% CI)	Level 3 variance (95% CI)	Variance of the linear predictors	% variance explained	DIC statistic
Null model	0.11	0.04	-	-	43,161
Research question (1)					
Option A *Area unemployment respondent*	0.16 (0.13-0.20)	0.05 (0.03-0.08)	0.85	19.5%	38,035
Option C *Area unemployment census*	0.07 (0.04-0.10)	0.03 (0.02-.05)	0.92	21.4%	37,056
Research question (2)					
Option B *Area IMD deciles*	0.08 (0.05-0.11)	0.02 (0.01-0.03)	0.93	21.6%	37,756
Option C *Area IMD score*	0.07 (0.04-0.10)	0.01 (0.01-0.02)	0.97	22.4%	37,688

While diagnostic statistics assess the multilevel model itself, and hence the quality of the prediction of whether an *individual* has an LLTI, synthetic estimates are concerned with predicting the prevalence of LLTIs at a *neighbourhood* or *small area* level. Figure 2 shows scatter plots of the synthetic estimates from each model against data from the 2011 Census. The fitted regression lines are positive in each case though the points in each graph are clearly scattered around the best-fit line.

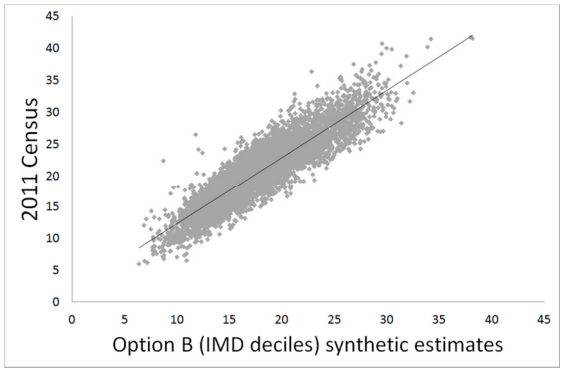
Figure 2 Comparing the synthetic estimates against the 2011 Census

Research question (1)





Research question (2)



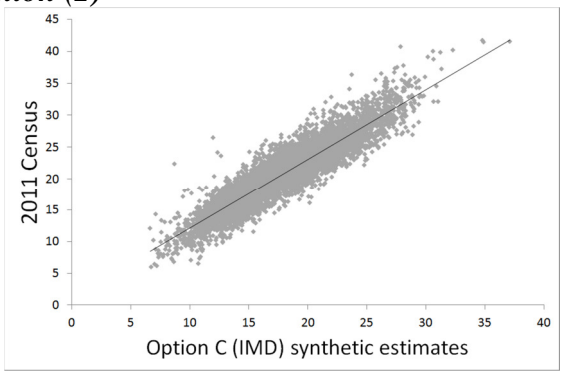


Table 4 explores the scatterplots more fully. The Spearman's rank correlation is least strong for Option A (area information based on aggregating survey responses) at r_s =0.64 compared with 0.85 for the synthetic estimates based on a multilevel model linking in census estimates of unemployment rates. This suggests that the synthetic estimates generated using Option C are a better reflection of the 2011 UK Census). With respect to the second research question, there is less to choose between the two sets of synthetic estimates, with Option C being slightly more strongly correlated with the census (0.92) than Option B (0.90).

Results based on the more stringent test of whether the synthetic estimates differ significantly from the line $x = y$ are less clear cut. In all but one case the parameters of the regression lines deviate from the $x = y$ line. Although it appears, for both research questions, that the synthetic estimates based on the fully geocoded version of the survey (Option C) are visually

a slightly closer match to the Census data, the regression test does not offer confirmation. In all models the general indication is that, on average, the synthetic estimates tend to be lower than the Census results. Only with the model using linked census data on unemployment does the gradient of the relationship between the estimates and the Census data indicate a close match.

Table 4 Comparing the synthetic estimates against the 2011 Census

	Spearman	Intercept	Lower CI	Upper CI	Contains zero?	Gradient	Lower CI	Upper CI	Contains one?
Research question (1)									
Option A *Unemployment (aggregated data)*	0.64*	4.69	4.21	5.17	no	0.87	0.84	0.89	no
Option C *Unemployment (linked census)*	0.85*	2.30	2.00	2.61	no	1.01	0.99	1.03	yes
Research question (2)									
Option B *IMD deciles*	0.90*	1.80	1.58	2.02	no	1.05	1.04	1.06	no
Option C *IMD score*	0.92*	1.13	0.92	1.34	no	1.10	1.08	1.11	no

Notes:
All correlations significant at the 0.01 level (two-tailed).
CIs represent 95 per cent confidence intervals.
.

5. Discussion and Conclusions

Despite the relative richness of the sources available to UK researchers, the data requirements for ML-SASE are restrictive. This paper has highlighted three broad categories of restrictions on researchers' choice of secondary data sources for ML-SASE. We have summarised how it is possible to generate multilevel synthetic estimates without knowing exactly where respondents are located in geographical space, but also highlighted how knowing where respondents live gives researchers considerable freedom in both defining the small areas and increasing the choice of area level explanatory variables. Instead of being forced to make pragmatic decisions based purely on the pre-existing availability of area data either already attached to the survey dataset and/or aggregated from individual responses, a researcher's choice of area level variables can be theoretically driven.

The four simple models generated as part of this paper have illustrated how having area data attached to the source survey (Option B) or ideally having the freedom to choose which area data are attached to the survey (Option C) can improve the quality of synthetic estimation models. Our results are less persuasive in terms of identifying improvements to the association between synthetic estimates and gold-standard direct estimates from Census data but we note that, in the majority of synthetic estimation exercises, such a gold standard would not be available.

The results demonstrate how being able to determine the nuances of the area level influences within a multilevel framework with either more accurate data and/or more precise data can improve point level synthetic estimates. We would contend that this improvement in accuracy, allied to the increased flexibility in choice of geography for the synthetic estimates suggests that survey datasets with spatial identifiers are preferable when selecting a source surveys to use as the basis for model development within a multilevel synthetic estimation framework.

While these results add to the literature on multilevel small area synthetic estimation, we acknowledge limitations to our analysis. We use relatively simple, though illustrative, multilevel models to generate synthetic estimates. We cannot be certain that the same conclusions would hold if more complex models had been built and recognise that increased model complexity can bring its own challenges, for example with respect to model convergence and model performance. Second, we cannot be sure that the same findings would necessarily hold if a different geography for the synthetic estimates had been selected and, indeed, different approaches to iterative estimation in multilevel software can lead to different outputs from the same underlying model (Browne and Draper, 2006). Third we have made comparisons that assume that census data provide a gold standard comparison. However, differences in the question wording between the Census and a chosen survey as well as the other issues such as question order and mode of questioning as well as census non-response mean that this may not necessarily be the case (Taylor, Twigg and Moon, 2014).

Nonetheless our paper has policy implications. It highlights the need for those responsible for commissioning large scale surveys to make available a geocoded version of their datasets. The 2010/11 sweep of the Crime Survey for England and Wales was, at the time of writing, available via the UK Data Service's special licence. This survey is ideal for multilevel synthetic estimation as it includes small area spatial codes. Some other surveys available

under special licence also meet this requirement. However, the spatial codes available under a special licence are often at a relatively coarse geography such as Local Authority. Data with small area geographical codes (such as postcodes or Super Output Areas) are generally only available to researchers from their institutional desktop in a virtual secure lab such as the UK Data Service's Secure Lab or from safe centre facilities. Micro safe settings or 'SafePods' at university/research institutes are likely to become increasingly common as a route to access secure data (Administrative Data Taskforce, 2013).

Unfortunately, not all large scale UK datasets have enhanced geocoded versions lodged in secure settings. For example, the Integrated Household Survey comprises a core suite of questions from other government surveys and currently represents the biggest pool of UK social data after the census. However, despite its obvious advantage in terms of both its size and topic coverage there is no Secure Lab version and the lowest geographical area for which data are available is Local Authorities. Secure laboratories represent a solution of considerable potential but to realise that potential they need to embrace full flexibility in terms of postcode level identifiers that will allow researchers to create their own geographies and not rely on those they are given.

This paper provides novel quantitative evidence testing our a priori expectation that surveys with spatial identifiers can result in more accurate synthetic estimates. We have showed that having access to geocoded survey datasets can have a positive effect on the specification of the model underpinning the synthetic estimates. Due to the increasing requirements for localised data both for policy and research purposes we therefore argue that it is imperative that access to survey datasets with spatial identifiers, such as postcodes and/or small area census geographies, must be maintained or ideally expanded.

Acknowledgements

This research was funded by the UK Economic and Social Research Council through its Secondary Data Research Initiative (grant reference ES/K003046/1). The authors acknowledge data access via the UK Data Service. The small area estimates referred to in this paper are available from http://reshare.ukdataservice.ac.uk/851972/.

References

Administrative Data Taskforce. (2013). Structuring the administrative data research network: report of the ADT technical group. Swindon: ESRC.

Alker, H. (1969). A typology of ecological fallacies. In M. Dogan & S. Rokkan (Eds.), *Quantitative ecological analysis in the social sciences* (pp. 69-86). London: Cambridge Press.

Bajekal, M., Scholes, S., Pickering, K., & Purdon, S. (2004). Synthetic estimation of healthy lifestyles indicators: Stage 1 report. London: National Centre for Social Research.

Browne, W. (2012). MCMC estimation in MLwiN version 2.26. Bristol: Centre for Multilevel Modelling, University of Bristol.

Browne, W., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*, 473-514.

CACI. (2009). ACORN user guide. London: CACI.

Chandra, H., Salvati, N., Chambers, R., & Tzavidis, N. (2012). Small area estimation under spatial nonstationarity. *Computational Statistics and Data Analysis, 56*, 2875-2888. doi: 10.1016/j.csda.2012.02.006

Dale, A. (2006). Quality issues with survey research. *International Journal of Social Research Methodology, 9*(2), 143-158. doi: 10.1080/13645570600595330

Dorling, D., & Thomas, B. (2004). *People and Places: a 2001 Census Atlas*. Bristol: Policy Press.

Duncan, C., Jones, K., & Moon, G. (1999). Smoking and deprivation: Are there neighbourhood effects? *Social Science and Medicine, 48*(4), 497-505.

Fitzpatrick, A., & Grant, C. (2011). 2010-11 British Crime Survey (England and Wales) Technical Report Volume I. London: TNS-BMRB.

Heady, P., Clarke, P., Brown, G., Ellis, K., Heasman, D., Hennell, S., . . . Mitchell, B. (2003). *Model-based small area estimation series number 2: Small area estimation project report*. London: ONS. Retrieved from http://www.statistics.gov.uk/methods_quality/downloads/small_area_est_report/saep1_Prelims&ch1&2_v2.pdf.

Henderson, M., Stansfeld, S., & Hotopf, M. (2013). Self-rated health and later receipt of work-related benefits: evidence from the 1970 British Cohort Study. *Psychological Medicine, 43*(8), 1755-1762.

Kass, R., Carlin, B., Gelman, A., & Neal, R. (1998). Markov Chain Monte Carlo in practice: A roundtable discussion. *The American Statistician, 52*(2), 93-100.

Marchetti, S., Tzavidis, N., & Pratesi, M. (2012). Non-parametric bootstrap mean squared error estimation for M-quantile estimators of small area averages, quantiles and poverty indicators. *Computational Statistics & Data Analysis, 56*(10), 2889-2902.

McLennan, D., Barnes, H., Noble, M., Davies, J., Garratt, E., & Dibben, C. (2011). *The English indices of deprivation 2010*. London: Department for Communities and Local Government Retrieved from http://www.communities.gov.uk/documents/statistics/pdf/1870718.pdf.

Molina, I., & Rao, J. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics, 38*(3), 369-385.

Norman, P., & Bambra, C. (2007). Incapacity or unemployment? The utility of an administrative data source as an updatable indicator of population health. *Population, Space and Place, 13*(5), 333-352. doi: 10.1002/psp.458

ONS. (2010). *Final recommended questions for the 2011 Census in England and Wales: health*. Titchfield: ONS. Retrieved from http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/2011-census-questionnaire-content/question-

and-content-recommendations-for-2011/final-recommended-questions-2011---health.pdf.
Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science, 28*(1), 40-68.
Pickering, K., Scholes, S., & Bajekal, M. (2004). Synthetic estimation of healthy lifestyles indicators: Stage 2 report. London: National Centre for Social Research.
Rafferty, A. (2009). Introduction to Complex Sample Design in UK Government Surveys Retrieved 18 June 2014, from http://esds.ac.uk/Government/docs/complexsampledesign.pdf.
Raftery, A., & Lewis, S. (1992). How many iterations in the Gibbs sampler? In J. Bernardo, J. Berger, A. Dawid & A. Smith (Eds.), *Bayesian statistics four* (pp. 763-773). Oxford: Oxford University Press.
Rasbash, J., Browne, W., Healy, M., Cameron, B., & Charlton, C. (2014). MLwiN version 2.30. , University of Bristol. University of Bristol: Centre for Multilevel Modelling.
Robinson, W. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*, 351-357.
Scarborough, P., Allender, S., Rayner, M., & 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 and Place, 15*, 596-605.
Scholes, S., Pickering, K., & Deverill, C. (2008). Healthy lifestyle behaviours: Model based estimates for middle layer super output areas and Local Authorities in England, 2003-2005: Stage 1 report. Leeds: The NHS Information Centre for health and social care.
Snijders, T., & Bosker, R. (2012). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling*: SAGE Publications.
Spiegelhalter, D., Best, N., Carlin, B., & A. van der Linde. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B - Statistical Methodology, 64*(4), 583-639. doi: DOI: 10.1111/1467-9868.00353
Subramanian, S., Jones, K., Kaddour, A., & Krieger, N. (2009). Revisiting Robinson: The perils of individualistic and ecologic fallacy. *International Journal of Epidemiology, 38*(2), 342-360.
The Countryside Agency, Department for Environment Food and Rural Affairs, Office of the Deputy Prime Minister, Office for National Statistics, & Welsh Assembly Government. (2004). *Rural and urban area classification 2004 an introductory guide*. London: Office for National Statistics. Retrieved from http://archive.defra.gov.uk/evidence/statistics/rural/documents/rural-defn/Rural_Urban_Introductory_Guide.pdf.
Taylor,J., Twigg, L., & Moon, G. (2014) The convergent validity of three surveys as alternative sources of health information to the 2011 UK census. *Social Science & Medicine*, *116*, 187-192.
Twigg, L., Moon, G., & Jones, K. (2000). Predicting small-area health-related behaviour: A comparison of smoking and drinking indicators. *Social Science and Medicine, 50*, 1109-1120.
Twigg, L., Moon, G., & Walker, S. (2004). *The smoking epidemic in England*. London: Health Development Agency. Retrieved from http://www.nice.org.uk/niceMedia/documents/smoking_epidemic.pdf.
UK Data Service. (2013). Introducing our data and services. Colchester: University of Essex.

Vickers, D., & Rees, P. (2007). Creating the UK National Statistics 2001 output area classification. *Journal of the Royal Statistical Society Series A - Statistics in Society, 170*, 379-403.

Whitworth, A. (2012). Sustaining evidence-based policing in an era of cuts: Estimating fear of crime at small area level in England. *Crime Prevention and Community Safety, 14*(1), 48-68.

Footnotes

[i] 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 1,000 and 3,000 can be amalgamated into larger Middle Layer Super Output Areas (MSOAs) each with a population of between 5,000 and 15,000.

[ii] Both the Crime Survey for England and Wales and the Health Survey for England employ complex sample designs. The confidence intervals calculated take into account the clustering and / or stratification using design factors (1.2 and 1.4 respectively). More information on complex survey designs can be found at Rafferty (2009).