ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2018)

A design-based approach to small area estimation using a semiparametric generalized linear mixed model

Hongjian Yu and Yueyan Wang,

University of California at Los Angeles, USA

Jean Opsomer,

Colorado State University, Fort Collins, USA

and Pan Wang and Ninez A. Ponce

University of California at Los Angeles, USA

[Received April 2016. Final revision December 2017]

Summary. In small area estimation, non-parametric models with penalized spline regression have been demonstrated to be a useful tool in creating granular area estimates to provide supplemental information where samples are few or non-existent. This study further examines the ability of a semiparametric generalized linear mixed model to produce conforming estimates for multiple area levels. A mosaic analogy is used to describe this process. A design-based jackknife method is employed for variance calculation.

Keywords: Logistic regression; Penalized spline; Small area estimation; Survey design

1. Introduction

Population health data at granular geographic levels are critical for describing health needs as well as designing and evaluating health programmes by health planners and policy makers. However, data from population-based health surveys usually lack sufficient geographic resolution for making estimates for local jurisdictions (Portnoy *et al.*, 2014). As population-based surveys are becoming more expensive, increasing survey sample sizes to yield community level estimates becomes less of an option. Small area estimation (SAE) techniques provide a solution to supplement survey data that do not provide reliable local estimates (Rao and Molina, 2015). SAEs from survey data have been used widely to generate official statistics in the USA (Mendez-Luck *et al.*, 2007; Yu *et al.*, 2007; Wang *et al.*, 2015; Nandram and Choi, 2002; Ghosh and Steorts, 2013; Nandram *et al.*, 2013; Jia *et al.*, 2004, 2006; Xie *et al.*, 2007; Schneider *et al.*, 2009; Srebotnjak *et al.*, 2010; Dwyer-Lindgren *et al.*, 2013, 2014; Linder *et al.*, 2013; Koh *et al.*, 2015; Berkowitz *et al.*, 2016; Pierannunzi *et al.*, 2016; Song *et al.*, 2016) and Europe (Fabrizi *et al.*, 2007; Giusti *et al.*, 2012; Longford *et al.*, 2012; Marchetti *et al.*, 2012). However, large-scale SAE projects based on survey data are not without challenges. A major challenge

Address for correspondence: Hongjian Yu, Center for Health Policy Research, University of California at Los Angeles, Suite 1550, 10960 Wilshire Boulevard, Los Angeles, CA 90024, USA.
E-mail: hyu@ucla.edu

 0964–1998/18/182000

is that areas of interest are so small that the samples are often sparse or non-existent. In the US setting, for example, it is unlikely that a state level survey would have a sufficient number of observations for every zip code that is of interest. Common SAE techniques, which have area-specific effects explicitly expressed in the models, may not perform well in such circumstances. Therefore, there is a need to combine survey data with other sources of data that are increasingly more detailed and abundant. It has been demonstrated in using the US Census Bureau's American Community Survey in several official statistics projects (Wang *et al.* (2015); `https://www.cdc.gov/500cities/`).

For data dissemination platforms that publish estimates at multiple geographical levels, another challenge is maintaining conformity across estimates that are reported at different levels of aggregation, i.e. when SAE estimates at various levels are aggregated to a larger level, such as counties or regions, the aggregated values should be statistically identical. The canonical approach to SAE would employ multiple models with random effects at respective geographic levels of interest. Although this approach is reasonable given the prediction targets at each level, the resulting predictions are not additive across scales. In practice, this might be solved by calibrating the small area estimates to higher level control totals by using a single model to generate estimates at multiple levels (Wang *et al.*, 2015). However, Wang *et al.* (2015) described only one approach tailored to a specific survey and did not provide an evaluation framework of alternative approaches, and thus it may limit the generalizability for addressing conformity across different surveys.

This study focuses more on the methodological details for the method proposed. Specifically, this study addresses the design consistency of the resulting small area estimate. A design consistent small area estimator converges to a true finite population quantity in probability under the randomization distribution. The study also addresses the seemingly contradictory combination of model-based estimation with design-based uncertainty assessment. In addition, this study also contributes to model selection for maintaining conformity of multiple levels of SAEs. A 'mosaic analogy' is illustrated as a guideline for such a purpose. We use a logistic regression model for demonstration given widespread use of the proportion of outcomes as local health indicators. The method can be generalized to other types of outcome.

In Section 2 we present the model and our method for SAE. The consistency of the SAE is discussed in detail in Section 3. Section 4 discusses variances for SAE. The estimation of small area estimates at multiple levels with conformity, and the elaboration of the 'mosaic' framework as a model selection guideline, is discussed through a demonstration in Sections 5 and 6. Throughout the following discussion we use a binary variable as the outcome of interest and logistic regression as the model function.

2. Model and small area estimation

The general approach to SAE that is discussed in this study is as follows. The model is built by using survey data; the estimated model coefficients are applied to a census level population data set containing the same set of independent variables; then predicted values are calculated for all individuals in the census. The population data have detailed geographic identifiers, which facilitated aggregating individual predicted values into small area level estimates. Population data have been utilized in SAE with unit level models in previous studies (Yu *et al.*, 2007; Mendez-Luck *et al.*, 2007).

The model under discussion is a unit level semiparametric logistic model of the form

$$\text{logit}(y_{ij}) = x_{ij}\beta + m(\xi) + w_i \tag{2.1}$$

where i is the index for areas and j is the index for individuals. It has three components: individual level fixed effects $\mathbf{x}$, a low rank P-spline with radial smoother $m(\xi)$ and, finally, an area level effect $\mathbf{w}$ for small areas of interest. The middle term ξ is a set of census tract level auxiliary contextual variables linked to data sets through the tract where the observation ij is nested. We used a P-spline to model this term because of its flexibility for unknown or otherwise complex data structure. Low rank means that the number of knots is far smaller than the number of data points and therefore reduces loss of parsimony compared with parametric models. This component can be visualized as a 'thin plate' smoothed over all the census tracts, serving as support for the small areas that have few or no observations. In addition, using a common form of minimization criterion, P-spline analysis can be converted to estimating the coefficients of a random-effect model (Ruppert *et al*., 2003; Opsomer *et al*., 2008). Therefore, the model can be fitted by using any commonly available software for mixed models.

Let s_i be the subsample for area i; following Rao and Molina (2015), a model-based small area proportion estimator has the form

$$\hat{\theta}_i = \frac{1}{N_i}\left(\sum_{j \in s_i} y_{ij} + \sum_{j \notin s_i} \hat{p}_{ij} \right). \tag{2.2}$$

The second term in parentheses is the summation of predicted probability for individuals who are not in the sample. Obviously, this estimator needs knowledge of the population. For that, we introduced the census data set that was mentioned above. The estimated model parameters are then applied to these census data. Let N_i and n_i be the population and sample sizes for area i and, assuming that $N_i \gg n_i$, we can approximate the above estimator by aggregating all the predicted values from the census data:

$$\hat{\theta}_i = \frac{1}{N_i} \sum_{ij} \hat{p}_{ij}. \tag{2.3}$$

3. Pseudo-maximum-likelihood estimator and design consistency of $\hat{\theta}_i$

For an easy illustration, we use a simplified model form. Let f be the inverse of the logit function for the model $p = f(x, \beta)$. Consider this model as generating the hypothetical superpopulation and the finite population as one of its realizations. The model parameter estimator β^* that is obtained by maximizing the census likelihood of the finite population would be consistent for β. When modelling survey data with an informative sampling design, Skinner (1989) developed the pseudo-maximum-likelihood (PML) method by following the ideas of Binder (1983). PML estimates the census likelihood by including sampling weights and maximizes it with the estimator $\hat{\beta}_{\mathrm{PML}}$. Under general conditions, $\hat{\beta}_{\mathrm{PML}}$ is consistent for β^*. Its asymptotic variance–covariance estimate is also unbiased. Given that β^* is model consistent for β, the PML estimator is jointly design consistent and model consistent.

PML is in general not connected to any optimization algorithm and most maximum likelihood estimation algorithms can be applied to include the weights and to maximize the PML function (Asparouhov, 2006).

For modelling in multistage surveys with unequal selection probabilities at stages, several works in the literature have established the need for models to account for data structure through multilevel models. This is because the units in the data are no longer independent; this non-independence needs to be reflected in the likelihood function for PML estimators to be consistent with superpopulation parameters. Although the effect on coefficient point estimates could be minimal, simulation studies show that variance estimates of PML estimates can

be severely biased if the original weights are directly applied, especially when cluster sizes at the individual unit level are small. When this happens, weight scaling is necessary to correct the bias (Pfeffermann *et al.*, 1998; Grilli and Pratesi, 2004; Asparouhov, 2006; Carle, 2009). Carle (2009) reviewed the software programs Mplus, MLwiN and GLLAMM that can perform proper modelling by either scaling the weights for users or allowing users to enter scaled weights. Recently the SAS procedure `GLIMMIX` has also added such capacity (Zhu, 2014). In the following discussion, we assume that the above issues are accounted for properly as needed.

Our descriptive population quantity of interest is the small area proportion

$$P_i = \frac{1}{N_i} \sum_{ij} y_{ij}.$$

An analytical form to approximate P_i under model f is

$$\theta_i = \frac{1}{N_i} \sum_{ij} p_{ij} = \frac{1}{N_i} \sum_{ij} f(x_{ij}, \beta^*).$$

The estimator for θ_i when plugging in the PML estimator for the model parameter becomes

$$\hat{\theta}_i = \frac{1}{N_i} \sum_{ij} \hat{p}_{ij} = \frac{1}{N_i} \sum_{ij} f(x_{ij}, \hat{\beta}_{\mathrm{PML}}). \tag{3.1}$$

Given model and design consistency, $\hat{\beta}_{\mathrm{PML}} \to \beta^*$ and, under the continuity condition of f, $\hat{\theta}_i$ converges to θ_i as the total sample size n increases.

If the model f holds for the population, then θ_i and P_i will converge as the total population size increases to ∞, i.e. $\theta_i \to P_i$ as $N \to \infty$ as well as $i = 1, \ldots, \infty$. The area index i also goes to ∞ because of random effects in the model. If the model does not hold, e.g. when one or more important predictors are absent, θ_i still has the interpretation of a population quantity (Pfeffermann, 1993). This offers a very meaningful benefit in SAE practice when models must be simplified for various technical or logistic reasons.

4. Variances

We define our method as design based because we assess uncertainty based on randomization in sample selection rather than model distribution. Its rationale can be demonstrated by following the decomposition in Pfeffermann (1993). Assuming that the model holds for the population and $N \gg n$, the consistency of $\hat{\theta}_i$ for P_i can be written as

$$\hat{\theta}_i - P_i = (\hat{\theta}_i - \theta_i) + (\theta_i - P_i) = O_p(n^{-1/2}) + O_p(N^{-1/2}) = O_p(n^{-1/2}).$$

Therefore the variance of $\hat{\theta}_i$ around P_i, which is denoted as $V(\hat{\theta}_i)$, is dominated by the randomization variance.

However, it may not be straightforward to obtain $V(\hat{\theta}_i)$. One way to estimate it is through parametric bootstrapping, similar to that in Opsomer *et al.* (2008). In this approach, we need to derive the estimate $\hat{\beta}_{\mathrm{PML}}$ and its covariance matrix $\hat{\Sigma}$ first, then to generate predicted model parameters from the distribution $\tilde{\beta} \sim \mathrm{MVN}(\hat{\beta}_{\mathrm{PML}} \hat{\Sigma})$ and finally to apply these $\tilde{\beta}s$ in estimator (3.1) for the distribution of $\hat{\theta}_i$.

For surveys with precalculated replicate weights, the process can be simpler. The replicate weights are intended to protect sample respondents' confidentiality, especially those from small primary sampling units. Keeping the fidelity of the original design, they are created through resampling from the original sample and then reweighted. The variance estimate is obtained

through the empirical distribution of the parameter estimates across these replicates (Rao and Wu, 1988; Stapleton, 2008). Use of these weights requires no knowledge of sample design (Rust and Rao, 1996; Wolter, 2007; Lohr, 2009). The California Health Interview Survey (CHIS) has 80 sets of replicate weights. Let θ denote the area parameter and $\hat{\theta}$ be its estimate derived from the model based on full sample weights. Let $\hat{\theta}_r$ be the estimate from the same model, but with rth replicate weights in the CHIS. The estimated variance is

$$\hat{V}(\hat{\theta}) = \sum_{r=1}^{80} \alpha_r (\hat{\theta}_r - \hat{\theta})^2$$

where α_r is the coefficient for replicate r, which equals 1 for all replicates in the CHIS.

5. Conforming estimates for uninsurance in Los Angeles County

Uninsurance estimates for adults aged 18–64 years in Los Angeles County at health district (HD) and zip code levels are used to illustrate our approach to generating conforming health estimates. Several candidate models are considered, and the process of choosing a better performing model is described. As the most populous county in California and in the USA with a diverse population of more than 10 million, Los Angeles County serves as a great example for examining health data at various geographic levels. Los Angeles County Department of Public Health divides the county into eight service planning areas (SPAs), which further subdivide into 26 HDs. HDs are spatially defined by grouping census tracts. These distinct jurisdictions enable the local public health department to develop and provide more relevant public health and clinical services targeted to the specific health needs of the residents in these socio-economically diverse areas (Los Angeles County Health Department, 2012).

US Postal Service zip codes are common proxies for referencing local geographies. The US Census Bureau approximates zip codes by using zip code tabulation areas (ZCTAs), which are based on census blocks. ZCTAs are much smaller than HDs, but ZCTAs do not neatly nest under HDs. ZCTAs, census tracts, HDs and SPAs serve as examples of a variety of geographic levels. In Los Angeles County, ZCTAs and census tracts are entities with comparable sizes, and HDs are much larger entities. Unlike ZCTAs, census tracts are nested within HDs which are further nested within SPAs, whereas ZCTAs cross boundaries of both census tracts and HDs.

5.1. Data source

5.1.1. Survey data: 2011–2012 California Health Interview Survey

The CHIS is the largest continuous state health survey in the USA. It employs a stratified random sample of approximately 20 000 households annually. In the CHIS, there are 56 geographic strata. A stratum is typically an individual county or group of small counties. The two largest counties, Los Angeles and San Diego, are further divided into eight (SPAs) and six (Health and Human Services Agency regions) strata respectively. In this study, the 2011–2012 sample of adults aged 18–64 years in Los Angeles County from the CHIS data is used. There are a total of 6218 individual adults residing in 281 ZCTAs represented in the sample. Los Angeles County has 294 ZCTAs, which means that the CHIS has no observations in 13 ZCTAs.

5.1.2. Census level population data: the Claritas data

The Claritas data set is a projected population data set provided by Claritas, which is a private marketing research firm. On the basis of estimates by the Census Bureau, its population

projections at the census block group levels utilize various sources that include trends in the US Postal Service deliverable address counts and in consumer counts from the Equifax total source database. This is important because, as time passes, the decennial census becomes a less accurate representation of the current population.

We augmented the Claritas data by building a predictive model for income-to-poverty ratios by using multinomial logistic regression in the CHIS, and then applied the model to Claritas data to produce a predicted income-to-poverty ratio for each observation. In subsequent estimation this predicted value was treated as fixed.

Finally, the resulting data set was adjusted to multiple CHIS weighting dimensions by using proportional iterative fitting so that it represented the population from which the CHIS sample was drawn.

5.1.3. Contextual data: American Community Survey

We downloaded the American Community Survey 2006–2010 5-year summary tables at the census tract level. The 236 variables were divided into 22 groups based on their content. Since the variables within each group are highly correlated, we conducted a principal component analysis for each group of variables to extract the information. The first principal component of each group accounted for 17–66% of variability within the group. To condense the information further to make it useful in modelling, these first principal components from each group were used as inputs for a secondary principal component analysis, resulting in the final principal components to be used as contextual variables in the model. Only the first two final principal components were retained (pc1 and pc2). They together accounted for 39% of the total variability among the first principal components of each variable group. The principal components were merged into both survey and population data by their census tracts and used as contextual variables.

5.2. Candidate models

Our test case is generating subcounty uninsurance rates in Los Angeles County. To ensure conformity of estimates, we examined several candidate models to find one that performs well at both ZCTA and HD levels.

Using dichotomous uninsurance status as the outcome, the unit level candidate model is the same as model (2.1). The three components are individual level fixed effects $\mathbf{X}$ for subjects' age, sex, poverty-to-income levels and race or ethnicity, a low rank P-spline with radial smoother to approximate unknown function $m(\xi)$ for census tract level two-dimensional continuous auxiliary contextual variables $\xi = (\text{pc1}, \text{pc2})$ from the American Community Survey and finally a random-area intercept $\mathbf{w}$ for small areas of interest. For the last term specifically, we define $\mathbf{u}$ and $\mathbf{v}$ as area effects at ZCTA and HD level respectively. Given that the CHIS has stratified random samples, sampling weights are incorporated in model fitting in their original forms for PML estimates and scaling is not needed. We first fitted a model with two levels of area effects $\mathbf{u}$ and $\mathbf{v}$ simultaneously, i.e. $\mathbf{w} = \mathbf{u} + \mathbf{v}$. In the mixed model framework, these two random intercepts are assumed to be independent. If we let i and j be the indices for $\mathbf{u}$ and $\mathbf{v}$, and let $\mathbf{k}$ be the index of the census tract, and $\mathbf{l}$ be an index for individuals in the place that is covered by ZCTA $\mathbf{l}$, HD j and census tract l simultaneously, then the model can be detailed as

$$\text{logit}(p_{ijkl}) = x_{ijkl}\beta + m(\xi_k) + u_i + v_j. \tag{5.1}$$

The resulting small area estimates from model (5.1) are then compared with those of two other models:

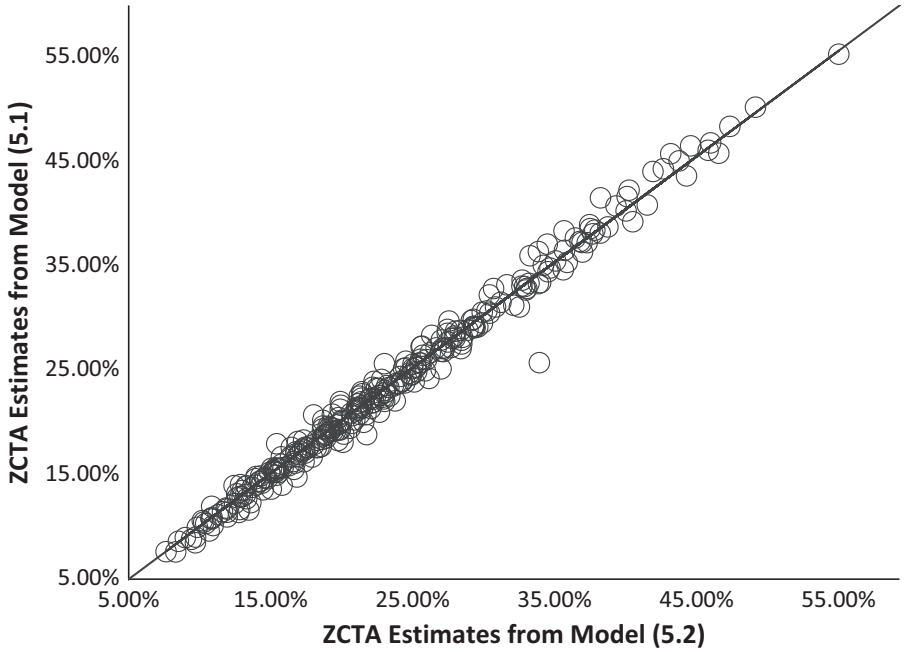

Fig. 1. Scatter plot of ZCTA estimates, model (5.1) *versus* model (5.2): model (5.1) and model (5.2) are both semiparametric models with a spline function of tract level auxiliary variables; model (5.1) has an area effect at both ZCTA and HD levels, whereas model (5.2) has only one area effect at ZCTA level

$$\text{logit}(p_{ijkl}) = x_{ijkl}\beta + m(\xi_k) + u_i, \tag{5.2}$$

$$\text{logit}(p_{ijkl}) = x_{ijkl}\beta + m(\xi_k) + v_j. \tag{5.3}$$

Model (5.2) and model (5.3) are both more parsimonious than model (5.1). Model (5.2) is the canonical model form for ZCTAs and model (5.3) is similarly that for HDs. For any ZCTA that has no observations in the CHIS, we substituted 0 as the predicted area effect for that ZCTA.

In knot specification for fitting *P*-splines, Ruppert (2002) recommended a default of 35–40 knots in general. We compared models with knots of 50, 35 and 30, all else equal. We used a standard space filling algorithm to place the knots for each number of knots (Royle and Nychka, 1998). Then we compared each set of area estimates. All three models produced similar estimates at both HD and ZCTA levels, at 30, 35 or 50 knots. We selected 30 knots for parsimony.

We used the `GLIMMIX` procedure from SAS 9.3, which accommodates the inclusion of weights in the model fit. It fits a generalized linear mixed model by using a linearization method. However, the resulting pseudolikelihoods were not appropriate for model comparison because the pseudodata that it fitted changes with each model change. For this reason, we did most model comparisons through plotting the point estimates and comparing their variances. `GLIMMIX` executes an automated space filling algorithm for fitting *P*-spline regressions. However, to apply the fitted model parameters to population data, we performed the algorithm by using SAS's `OPTEX` procedure in the CHIS to derive knot locations (or co-ordinates). They were then used to calculate the design matrix for spline regression, which was linked to the population data. For details on these procedures, refer to SAS Institute (2013).

To demonstrate the utility of linked American Community Survey contextual variables we also compared model (5.2) with model (5.4) with **u** only. Both can be considered as canonical SAE models in which area-specific random effects are explicitly expressed in the model:

$$\text{logit}(p_{ijl}) = x_{ijl}\beta + u_i. \tag{5.4}$$

Note that SAEs generated from model (5.4) do not have the 'thin plate' support that is forged by *P*-splines.

5.3. Multilevel area estimation

In this section, we examine whether one relatively simpler model can generate satisfactory estimates for multiple levels of area parameters.

ZCTA estimates from models (5.1) and (5.2) were compared in a scatter plot in Fig. 1. If these two sets of estimates match exactly, all data points should lie on the 45° line. Fig. 1 shows that the two sets of estimates closely line up and indicates that including the HD effect in addition to the ZCTA effect contributed little extra information to the estimates at the ZCTA level. The correlation between the two sets of estimates is 0.99.

To visualize the effect of census tract auxiliary variables on ZCTA estimation, Fig. 2 shows a comparison between model (5.2) and model (5.4). The estimates from model (5.4) for ZCTAs with no observations depart from the 45° line and turn sideways. This indicates that relying on survey data alone to build a model is insufficient to preserve the local variations for the estimates

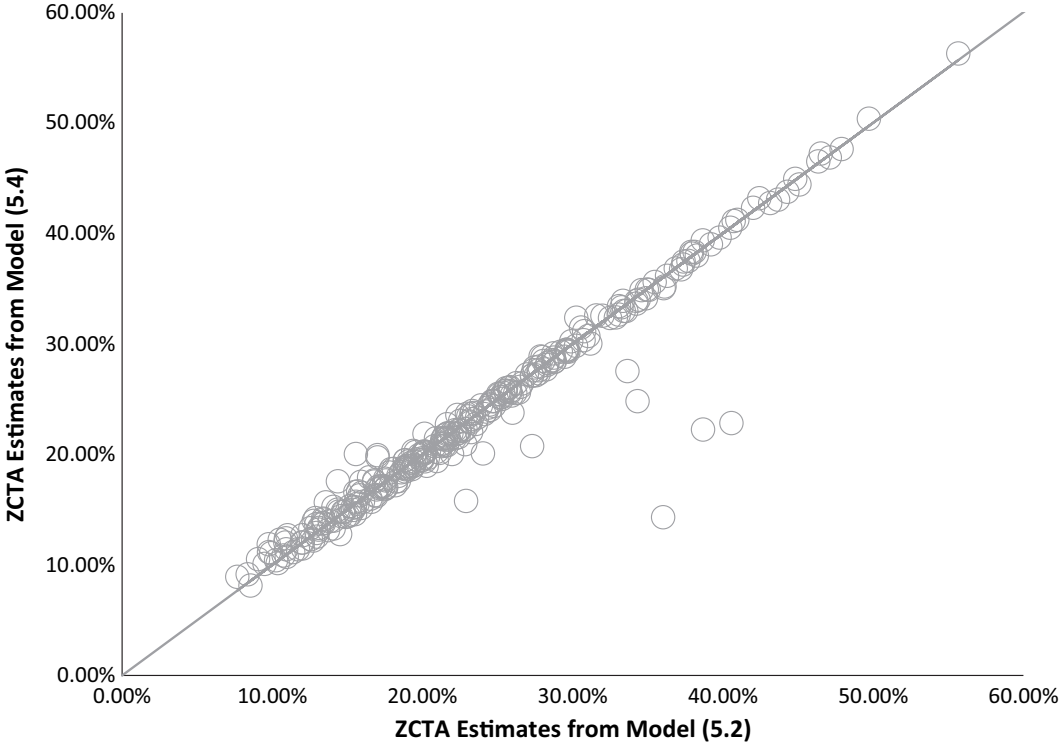

Fig. 2. Scatter plot of ZCTA estimates, model (5.2) *versus* model (5.4): model (5.2) is a semiparametric model with the spline term and an area effect at ZCTA level; model (5.4) is the canonical SAE with only ZCTA level effect but no tract level spline term

at the level of ZCTAs. When the fitted model is applied to the population data, the only source of variation among these estimates is from fixed individual predictors X in the population data. A similar pattern was also observed in Opsomer *et al*. (2008).

For estimation at HD level, both model (5.1) and model (5.3) include HD area level random effects. The estimates from these models together with those from model (5.2) are compared in Fig. 3. Although the estimates from model (5.1) are similar to those of (5.3), model (5.2) performs equally well, even though model (5.2) excludes HD area effects. Table 1 shows high

Table 1. Correlations between estimates from three semiparametric models, all with spline function at tract level but different area effects

	Model (5.3) HD only	*Model (5.1) HD + ZCTA*	*Model (5.2) ZCTA only*
Model (5.3) HD only	1		
Model (5.1) HD + ZCTA	0.98	1	
Model (5.2) ZCTA only	0.96	0.98	1

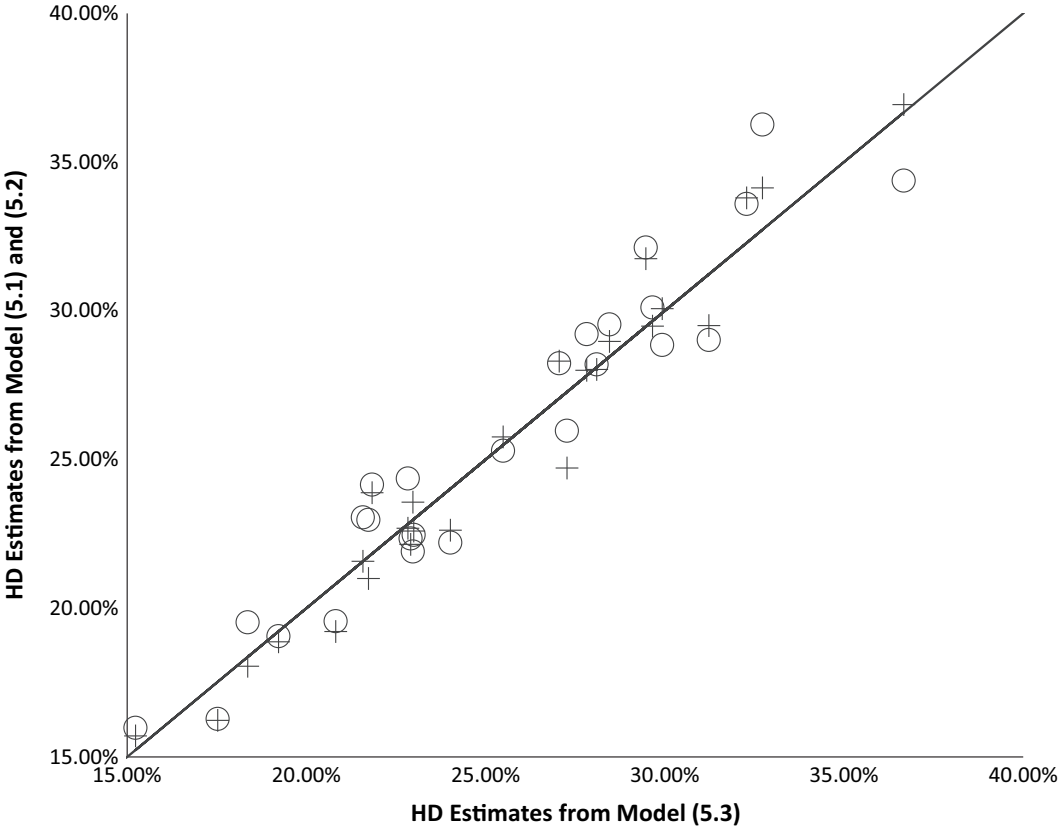

Fig. 3. Scatter plot of HD estimates from three models: model (5.1) (+) with both ZCTA and HD area effect, model (5.2) (○) with ZCTA area effect only and model (5.3) with HD area effect only (all three models were semiparametric models sharing a spline function of census tract auxiliary variables)

correlations between the three sets of estimates. Considering both Fig. 1 and Fig. 3, it indicates that model (5.2) with both the spline and **u** (ZCTA area effect) performs sufficiently well for both ZCTA and HD level estimates.

However, this is not so in reverse; model (5.3) with **v** (HD area effect) only performs poorly in ZCTA estimation. Fig. 4 displays the comparison of ZCTA estimates from model (5.2) against those from model (5.3). It clearly shows the deficiency of model (5.3) in ZCTA estimation. The correlation between the two is 0.71.

This invokes an analogy with a mosaic, in which details in small pieces (at ZCTA level) will make an accurate bigger picture (at HD level). The mosaic implication is a simplified model for these two levels of estimation, even though the ZCTA and HD models are not hierarchical. To look into this further, we examined a fifth model (5.5) in which **w** is the area effect at census tract level.

Since the HDs are aggregations of census tracts, the latter are nested within the former. HD estimates from model (5.5) are similar to the estimates from models (5.1)–(5.3), as depicted in Fig. 3. This result also supports the mosaic analogy. For clarity in the paper, we did not include the HD estimates in the illustrations. The correlations between estimates from model (5) and those from model (5.2) and model (5.3) are both 0.95.

In contrast, model (5.5) performs poorly for ZCTA level estimation, as shown in Fig. 5. Since the size difference between zip code and census tract is not as great as that between ZCTA and

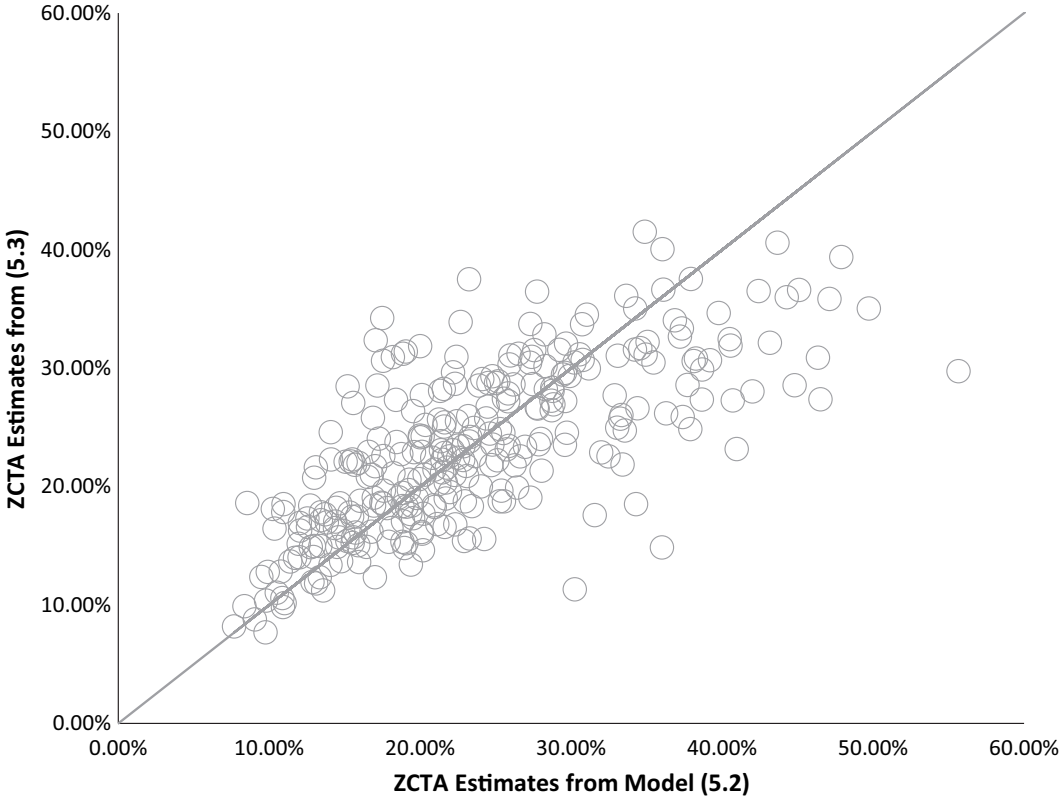

Fig. 4. Scatter plot of ZCTA estimates, model (5.2) *versus* model (5.3): model (5.2) and model (5.3) are both semiparametric models with a spline function of tract level auxiliary variables; model (5.2) has an area effect at ZCTA level, whereas model (5.3) has an area effect at HD level

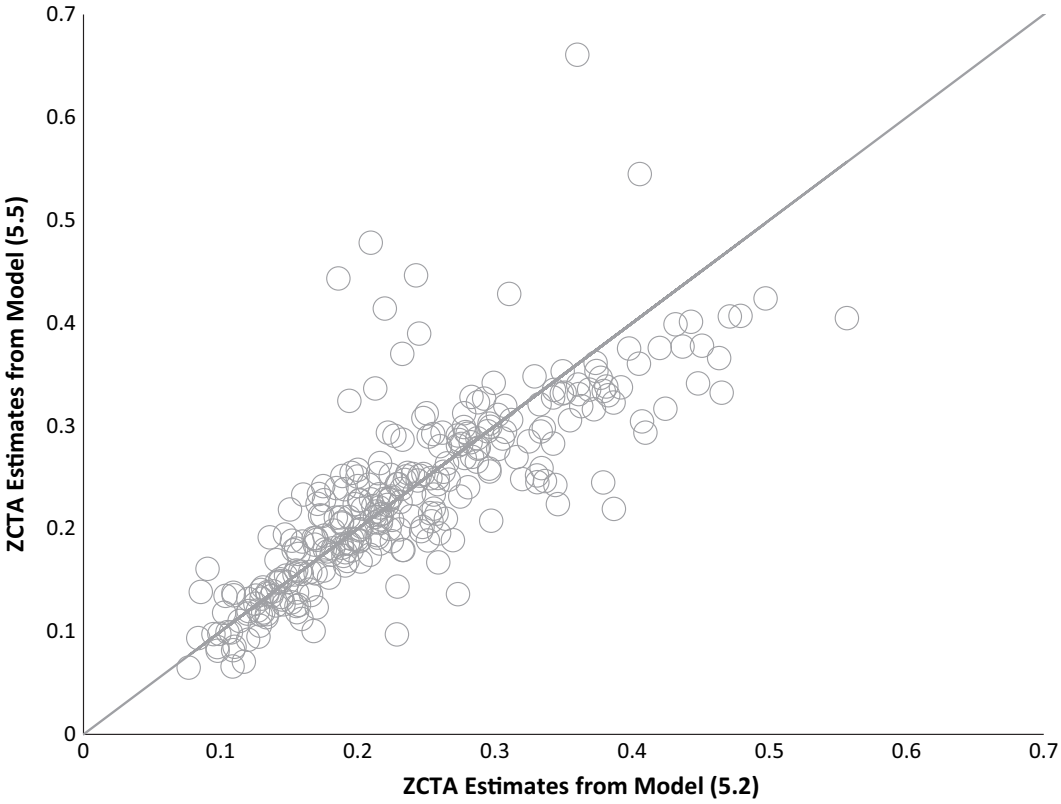

Fig. 5. Scatter plot of ZCTA estimates, model (5.2) *versus* model (5.5): model (5.2) and model (5.5) are both semiparametric models with a spline function of tract level auxiliary variables; model (5.2) has an area effect at ZCTA level, whereas model (5.5) has an area effect at census tract level

HD, one could think that tracts were not sufficiently detailed to serve well as mosaic pieces for ZCTAs. The correlation between the two sets of estimates in Fig. 5 is 0.80.

6. Variance estimation by using survey replicate weights

Fig. 6 shows variance estimates of ZCTA estimations from model (5.2) *versus* those from model (5.1). Once again, it shows that the HD effect has hardly any effect on the ZCTA variance estimates. Fig. 7 shows variance estimates of HD estimations from the two models. For supplementary comparison, we also added variance estimates from model (5.3): the model with HD area effect. Although it is expected that model (5.2) produces smaller variances than model (5.1) because it is a simpler model than model (5.1), it is interesting that its variances are even smaller than those for model (5.3) for most HDs, as model (5.3) is the canonical model for HDs, in model-based terms, whereas model (5.2) is not.

The results in both Section 2 and Section 3, especially those for HD estimates, point to model (5.2). All things considered, model (5.2) performs as equally well as or better than all the other models that were examined although being parsimonious.

Finally, we generated estimates for the Los Angeles SPAs from model (5.2) and compared them with CHIS direct estimates that are stable. Fig. 8 shows that the modelled estimates match

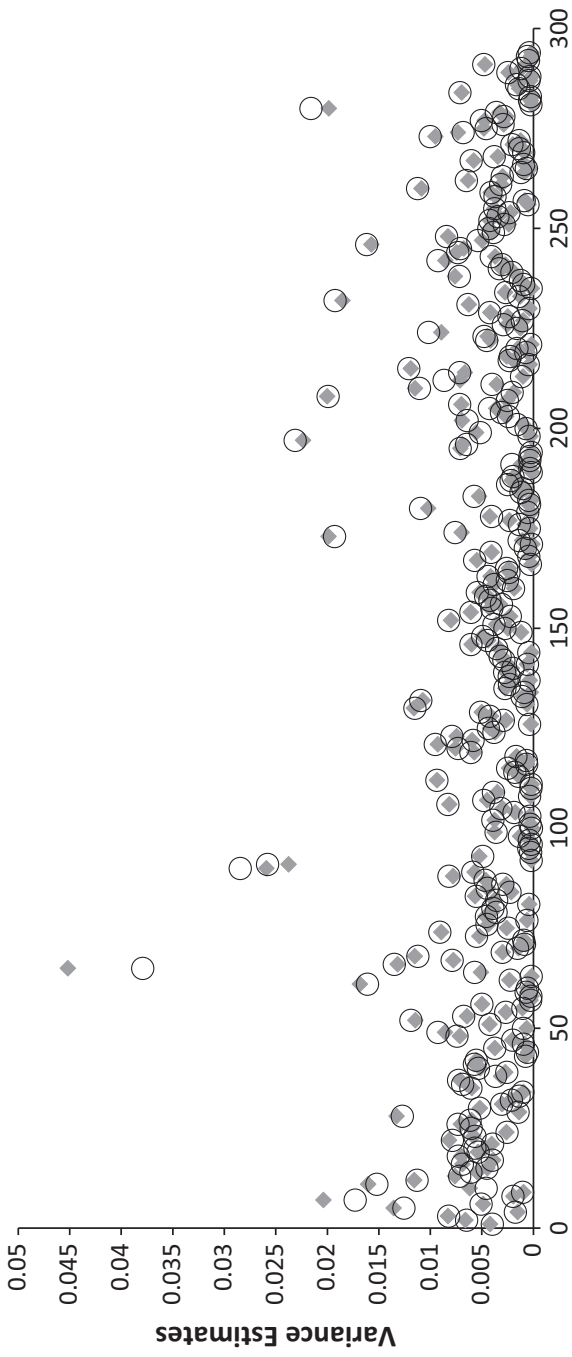

Fig. 6. Variance estimates of ZCTA estimations from model (5.1) (○) and model (5.2) (◆): model (5.1) and model (5.2) are both semiparametric models with a spline function of tract level auxiliary variables; model (5.1) has area effect at both ZCTA and HD levels, whereas model (5.2) has only one area effect at ZCTA level

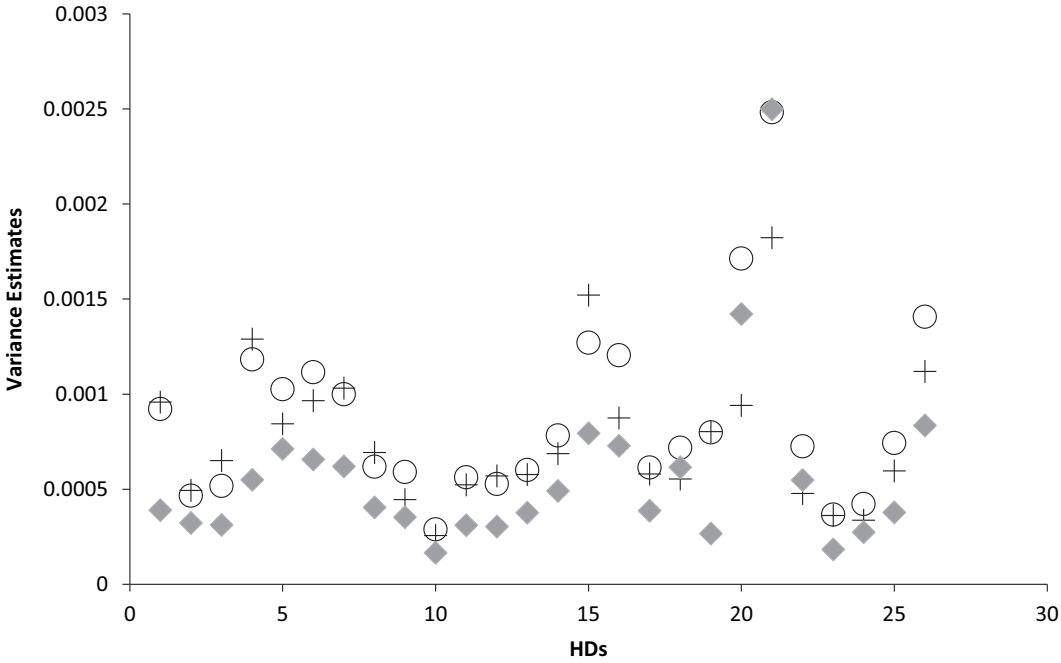

Fig. 7. Variance estimates of HD estimations from three models: model (5.1) (○) with both ZCTA and HD area effect, model (5.2) (◆) with ZCTA area effect only and model (5.3) (+) with HD area effect only (all three models were semiparametric models sharing a spline function of census tract auxiliary variables)

very well with the direct estimates although the SPA effect was not specified in the model. This can serve as another validation of the mosaic analogy. It also implies the consistency of the estimator $\hat{\theta}_i$ (3.1). In addition, we see a vast improvement in precision by using modelled estimates as opposed to direct estimates.

7. Conclusion and discussion

We presented a method that is capable of tackling some common challenges that practitioner statisticians may face in large-scale production of small area estimates. The method encompasses survey data and census tract auxiliary variables from the American Community Survey for conforming SAEs across various geographic levels. Using the population data enabled an easier approach to consistent SAE. The use of non-parametric functions of community level contextual variables captures the association between local socio-economic status and the outcome in a flexible and efficient manner. In dealing with multiple levels of SAE, a mosaic analogy served as a guideline in model construction for estimators at various levels: the area effects at the most granular level are most critical for the model.

Employing population data is beneficial in multiple ways. They provide a means to apply unit level models with non-linear link functions such as logistic models in SAE as demonstrated. They offer a simpler approach to design consistency of SAE by just focusing on the consistent model parameter estimators. They provide flexibility to generate estimates for areas that may not coincide with the survey design, such as zip codes or legislative districts. They also enhance the ability to estimate for the areas that have no observations in a survey. However, they have shortcomings. A population data set may have limited predictors which in turn limits models'

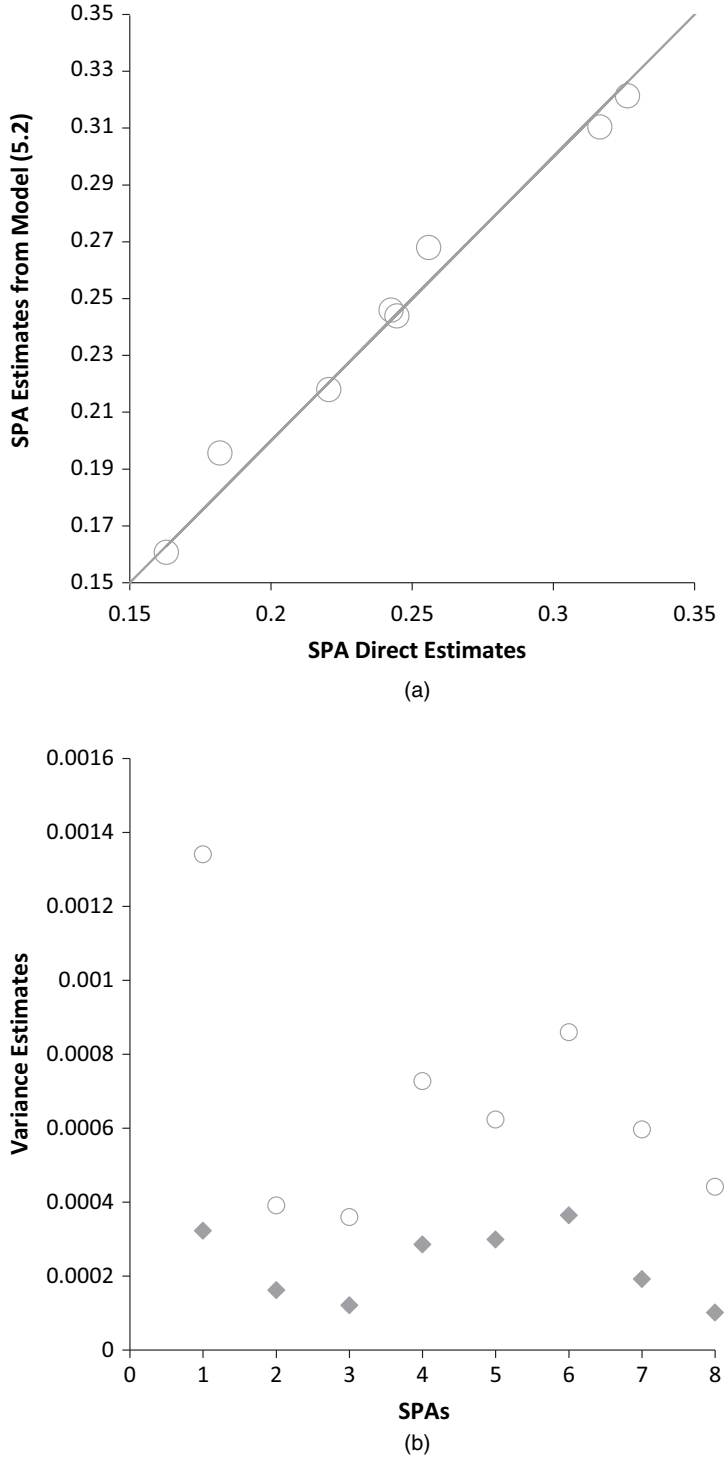

Fig. 8. Scatter plot of (a) SPA estimates and (b) variances, model (5.2) (○) *versus* CHIS direct estimates (○): model (5.2) is a semiparametric model with ZCTA area effect and the spline term

ability to employ rich information in a survey. They can also be difficult and expensive to acquire and construct, especially between decennial census years.

An important issue in unit level modelling is that design informativeness needs to be accounted for when fitting models. This can be incorporated by performing survey-weighted estimation and design-based inference. This leads to valid SAE either by assuming that the resulting estimates are jointly consistent with respect to the design and the model or, more conservatively, that the estimation targets are the descriptive population quantities rather than the underlying model characteristics. Under either setting, the resulting methods are design based in that we approximate the uncertainty by focusing the variance from randomization alone rather than from joint randomization and model distribution. This also relieves the practitioner of the heavy burden of model assumption validation, which in many cases is infeasible. Another potential advantage of a design-based method is the ease of adopting new modelling techniques such as machine learning where an explicit likelihood function may not be available. Applying replicate weights in such cases is relatively straightforward. Nevertheless, validations such as comparing with direct estimates, e.g. Fig. 8, or with data from external sources are important to evaluate and ensure the quality of estimates.

In the majority of applications of SAE, a purely model-based approach is preferred, either for tractability or because the design is truly non-informative. In fact, the latter assumption is often reasonable for area level models, but much more problematic for unit level models that are similar to those considered in this study. By fitting a small area model with survey weights, we can use the informative utility of the design, which is a clear advantage of the approach. Another advantage is that variance estimation is straightforward: it follows from standard design-based asymptotic theory. In addition, in cases where replication weights are provided as part of the data set, variance estimation is even more straightforward.

Acknowledgements

Financial support for this research was provided by an 'AskCHIS neighborhood edition' grant from a Kaiser Permanente community benefits programme and the California Wellness Foundation.

We are grateful for insightful comments, suggestions and criticism from the reviewers. We are also grateful to Carl Ganz for the support that we received in revision. We shall always be grateful for the leadership and wisdom of the late Dr E. Richard Brown, who developed the CHIS and provided great support for the development of small area estimates.

References

Asparouhov, T. (2006) General multi-level modeling with sampling weights. *Communs Statist. Theory Meth.*, **35**, 439–460.

Berkowitz, Z., Zhang, X., Richards, T. B., Peipins, L., Henley, S. J. and Holt, J. (2016) Multilevel small-area estimation of multiple cigarette smoking status categories using the 2012 behavioral risk factor surveillance system. *Cancer Epidem. Prevn Biomark.*, **25**, 1402–1410.

Binder, D. A. (1983) On the variances of asymptotically normal estimators from complex surveys. *Int. Statist. Rev.*, **51**, 279–292.

Carle, A. C. (2009) Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Med. Res. Methodol.*, **9**, no. 1, article 49.

Dwyer-Lindgren, L., Freedman, G., Engell, R. E., Fleming, T. D., Lim, S. S., Murray, C. J. and Mokdad, A. H. (2013) Prevalence of physical activity and obesity in US counties, 2001–2011: a road map for action. *Popln Hlth Metr.*, **11**, no. 1, article 7.

Dwyer-Lindgren, L., Mokdad, A. H., Srebotnjak, T., Flaxman, A. D., Hansen, G. M. and Murray, C. J. (2014) Cigarette smoking prevalence in US counties: 1996-2012. *Popln Hlth Metr.*, **12**, no. 1, article 5.

Fabrizi, E., Ferrante, M. R. and Pacei, S. (2007) Small area estimation of average household income based on unit level models for panel data. *Surv. Methodol.*, **33**, 187–198.
Ghosh, M. and Steorts, R. C. (2013) Two-stage benchmarking as applied to small area estimation. *Test*, **22**, 670–687.
Giusti, C., Marchetti, S., Pratesi, M. and Salvati, N. (2012) Robust small area estimation and oversampling in the estimation of poverty indicators. *Surv. Res. Meth.*, **6**, no. 3, 155–163.
Grilli, L. and Pratesi, M. (2004) Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. *Statist. Can.*, **30**, 93–103
Jia, H., Link, M., Holt, J., Mokdad, A. H., Li, L. and Levy, P. S. (2006) Monitoring county-level vaccination coverage during the 2004–2005 influenza season. *Am. J. Prev. Med.*, **31**, 275–280.
Jia, H., Muennig, P. and Borawski, E. (2004) Comparison of small-area analysis techniques for estimating county-level outcomes. *Am. J. Prev. Med.*, **26**, 453–460.
Koh, K., Grady, S. C. and Vojnovic, I. (2015) Using simulated data to investigate the spatial patterns of obesity prevalence at the census tract level in metropolitan Detroit. *Appl. Geog.*, **62**, 19–28.
Linder, J. A., Rigotti, N. A., Brawarsky, P., Kontos, E. Z., Park, E. R., Klinger, E. V., Marinacci, L. and Haas, J. S. (2013) Use of practice-based research network data to measure neighborhood smoking prevalence. *Prev. Chron. Dis.*, **10**, article E84.
Lohr, S. (2009) *Sampling: Design and Analysis*, 2nd edn. Boston: Cengage Learning.
Longford, N. T., Pittau, M. G., Zelli, R. and Massari, R. (2012) Poverty and inequality in European regions. *J. Appl. Statist.*, **39**, 1557–1576.
Los Angeles County Health Department (2012) Los Angeles County GIS data portal. (Available from `http://egis3.lacounty.gov/dataportal/2012/03/01/health-districts-hd-2012`.)
Marchetti, S., Tzavidis, N. and Pratesi, M. (2012) Non-parametric bootstrap mean squared error estimation for M-quantile estimators of small area averages, quantiles and poverty indicators. *Computnl Statist. Data Anal.*, **56**, 2889–2902.
Mendez-Luck, C. A., Yu, H., Meng, Y. Y., Jhawar, M. and Wallace, S. P. (2007) Estimating health conditions for small areas: asthma symptom prevalence for state legislative districts. *Hlth Serv. Res.*, **42**, 2389–2409.
Nandram, B., Bhatta, D., Bhadra, D. and Shen, G. (2013) Bayesian predictive inference of a finite population proportion under selection bias. *Statist. Methodol.*, **11**, 1–21.
Nandram, B. and Choi, J. W. (2002) A Bayesian analysis of a proportion under non-ignorable non-response. *Statist. Med.*, **21**, 1189–1212.
Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G. and Breidt, F. J. (2008) Non-parametric small area estimation using penalized spline regression. *J. R. Statist. Soc.* B, **70**, 265–286.
Pfeffermann, D. (1993) The role of sampling weights when modeling survey data. *Int. Statist. Rev.*, **61**, 317–337.
Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H. and Rasbash, J. (1998) Weighting for unequal selection probabilities in multilevel models. *J. R. Statist. Soc.* B, **60**, 23–40.
Pierannunzi, C., Xu, F., Wallace, R. C., Garvin, W., Greenlund, K. J., Bartoli, W., Ford, D., Eke, P. and Town, G. M. (2016) A methodological approach to small area estimation for the behavioral risk factor surveillance system. *Prev. Chron. Dis.*, **13**, article E91.
Portnoy, B., Lee, S. J. C., Kincheloe, J., Breen, N., Olson, J. L., McCormally, J. and Brown, E. R. (2014) Independent state health surveys: responding to the need for local population health data. *J. Publ. Hlth Mangmnt Pract.*, **20**, no. 5, E21–E33.
Rao, J. and Molina, I. (2015) *Small Area Estimation*, 2nd edn. Hoboken: Wiley.
Rao, J. N. K. and Wu, C. F. J. (1988) Resampling inference with complex survey data. *J. Am. Statist. Ass.*, **83**, 231–241.
Royle, J. A. and Nychka, D. (1998) An algorithm for the construction of spatial coverage designs with implementation in SPLUS. *Comput. Geosci.*, **24**, 479–488.
Ruppert, D. (2002) Selecting the number of knots for penalized splines. *J. Computnl Graph. Statist.*, **11**, 737–757.
Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression*. Cambridge: Cambridge University Press.
Rust, K. F. and Rao, J. N. (1996) Variance estimation for complex survey data using replicate methods. *Statist. Meth. Med. Res.*, **5**, 283–310.
SAS Institute (2013) *SAS/STAT 13.1 User's Guide: the Glimmix Procedure*. Cary: SAS Institute.
Schneider, K. L., Lapane, K. L., Clark, M. A. and Rakowski, W. (2009) Using small-area estimation to describe county-level disparities in mammography. *Prev. Chron. Dis.*, **6**, no. 4, article A125.
Skinner, C. J. (1989) Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys* (eds C. J. Skinner, D. Holt and T. M. F. Smith), pp. 59–87. Chichester: Wiley.
Song, L., Mercer, L., Wakefield, J., Laurent, A. and Solet, D. (2016) Using small-area estimation to calculate the prevalence of smoking by subcounty geographic areas in King County, Washington, behavioral risk factor surveillance system, 2009–2013. *Prev. Chron. Dis.*, **13**, article E59.
Srebotnjak, T., Mokdad, A. H. and Murray, C. J. (2010) A novel framework for validating and applying standardized small area measurement strategies. *Popln Hlth Metr.*, **8**, no. 1, article 26.

Stapleton, L. M. (2008) Variance estimation using replication methods in structural equation modeling with complex sample data. *Struct. Equn Modlng*, **15**, 183–210.

Wang, Y., Ponce, N. A., Wang, P., Opsomer, J. D. and Yu, H. (2015) Generating health estimates by zip code: a semiparametric small area estimation approach using the California Health Interview Survey. *Am. J. Publ. Hlth*, **105**, 2534–2540.

Wolter, K. (2007) *Introduction to Variance Estimation*. Chicago: Springer Science and Business Media.

Xie, D., Raghunathan, T. E. and Lepkowski, J. M. (2007) Estimation of the proportion of overweight individuals in small areas—a robust extension of the Fay–Herriot model. *Statist. Med.*, **26**, 2699–2715.

Yu, H., Meng, Y.-Y., Mendez-Luck, C. A., Jhawar, M. and Wallace, S. P. (2007) Small-area estimation of health insurance coverage for California legislative districts. *Am. J. Publ. Hlth*, **97**, 731–737.

Zhu, M. (2014) Analyzing multilevel models with the GLIMMIX procedure. In *Proc. SAS Global Forum 2014 Conf*. Cary: SAS Institute. (Available from `http://support.sas.com/resources/papers/proceedings14/SAS026-2014`.)