Development and evaluation of an optimal composite estimator in spatial microsimulation small area estimation

Revised Manuscript

Abstract

A range of data are of geographic interest but are not available at small area level from existing data sources. Small area estimation (SAE) offers techniques to estimate population parameters of target variables to detailed scales based on relationships between those target variables and relevant auxiliary variables. The resulting indirect small area estimate can deliver a lower mean squared error compared to its direct survey estimate given that variance can be reduced markedly even if bias increases. Spatial microsimulation SAE approaches are widely utilised but only beginning to engage with the potential of composite estimators that use a weighted combination of indirect and direct estimators to reduce further the mean squared error of the small area estimate compared to an indirect SAE estimator alone. This paper advances these approaches by constructing for the first time in the microsimulation literature an optimal composite estimator for such SAE approaches in which the combining weight is calculated from the mean squared errors of the two estimators, thus optimising the reduction in MSE of the resulting small area estimates. This optimal composite estimator is demonstrated and evaluated in a model-based simulation study and application based on real data.

Keywords: Small area estimation; calibration; expansion estimator; synthetic estimator; variance; composite estimation.

1. Introduction

Many social phenomena of interest to policy makers and scholars are known to be spatially heterogeneous, desired to be measured and understood at detailed spatial scales, but frequently unavailable at small area level from existing Census, administrative or commercial data sources – income, wellbeing, diet and exercise, attitudes, and so on. Existing large-scale sample survey data often do contain variables related to those phenomena but are not designed to be representative at a small geographical level and are expensive to conduct (Moretti et al., 2019; Buil-Gil et al., 2019). Therefore, the issue of unplanned domains arises where the problem of small or zero sample sizes exists at the small area level of interest. In this context direct design-based estimators such as the well-known Horvitz-Thompson estimator (Horvitz and Thompson, 1952) are limited given that for area with small sample sizes they provide large variability in estimates and that and in small areas with zero sample size they cannot be used (Rao and Molina, 2015).

In response, indirect small area estimation (SAE) techniques have been used to produce small area estimates where direct estimators are either not viable or unreliable. SAE offers techniques to estimate target parameters to detailed scales based on relationships between target variables and relevant covariates and their application to the same covariates at small area level. Rao and Molina (2015) provide a helpful review. SAE methods can be classified into two broad groups: spatial microsimulation and regression-based approaches (Whitworth, 2013; Rahman and Harding, 2017). Within regression-based approaches a range of modelling strategies have been adopted including ecological (Ipsos MORI, 2015), univariate mixed-effect (Battese et al, 1988), multivariate mixed-effect (Datta et al, 1999), M-Quantile (Marchetti et al, 2012) and Bayesian (Maiti, 2005). Three main spatial microsimulation approaches exist based on the

optimal reweighting of all survey respondents to the target small area profile – iterative proportional fitting (IPF) (Ballas et al, 2005) and generalized regression (GREGWT) (Singh and Mohl, 1996) and the combinatorial optimisation approach by Williamson et al (1998). Similarities exist between these approaches and it is indeed possible to consider them simply as alternative forms of either integer or non-integer reweighting approaches.

Irrespective of the SAE approach adopted all have the potential to counter the two fatal limitations of direct survey estimation to small area level – to deliver small area estimates in contexts with zero survey sample size and to reduce markedly the variance around the indirect small area estimates compared to the direct small area estimates. However, as with all synthetic estimators this process introduces bias into the indirect small area estimates, in contrast to the unbiased direct estimates (Berg and Fuller, 2009; Griffiths, 1996).

A natural way to tackle the trade-off between bias and variance in small area estimators is to construct weighted composite estimators between an unbiased (but with larger variability) direct estimator and a biased (but with smaller variability) synthetic estimator. These offer the potential to reduce further the mean squared error (MSE) of the small area estimate – the key metric of estimate quality that takes into account both bias and variance – compared to an indirect SAE estimator alone.

Such composite estimators are relatively widely used in regression-based SAE approaches (Rao and Molina, 2015). Within the popular spatial microsimulation SAE approaches, however, understanding and practice around the potential benefits of composite estimators is largely absent. Moretti and Whitworth (2019b) recently presented and evaluated a sample-size-dependent composite estimator, first introduced by Drew et al. (1982), into a spatial microsimulation SAE setting. In a sample-size-dependent approach, for each target small area the weighting attached to the direct and indirect sides respectively of the composite estimator are derived based on the sample size in that small area: as the small area sample size increases the weighting attached to the direct survey estimator relative to the indirect spatial microsimulation SAE estimator increases as the variability around that direct estimator becomes smaller, and vice versa. This was an important first attempt in the development of composite small area estimation within spatial microsimulation approaches. However, sample-size-dependent composite estimators are limited in that they do not take into account the size of the between-area variability relative to the within-area variability (Drew et al., 1982; Rao and Molina, 2015). As such, sample-size-dependent estimators are not well suited to contexts with high levels of between-area variation given that they are unable to capture this variation.

In contrast, optimal composite estimators can be used to overcome these limitations of neglecting the local variation with sample-size-dependent approaches. In addition, optimal composite estimators helpfully base the weighting attached to the direct and indirect sides of the composite estimator on the key metric of real interest to optimising the balance between bias and variance in the final small area estimates – the mean squared error. This is because in an optimal composite approach that key weighting between the direct and indirect sides of the composite estimator is based not on the survey sample size of each target area – which acts as an imperfect proxy for the real estimation interest in minimisation of the MSE – but instead explicitly on the actual MSEs of those direct and indirect estimators themselves (Rao and Molina, 2015; Schaible, 1978).

This article develops the first presentation and evaluation of a small area optimal composite estimator in the literature of spatial microsimulation framework. The focus in this article is on

the iterative proportional fitting (IPF) spatial microsimulation approach, though the principles are applicable more broadly across the different spatial microsimulation SAE methods. The remainder of the article is structured as follows. In Section 2 the notation used and direct survey approaches are briefly outlined before turning to the exposition of the IPF approach to SAE. In Section 3 the composite estimators are described beginning with the sample-size-dependent composite estimator and next the optimal composite estimator. Section 4 provides the results of the simulation study evaluating the comparative performance of the optimal and sample-size-dependent composite estimators in terms of the key MSE performance metric that takes into account both the bias and the variance of the resulting estimates. Section 5 offers an applied application based on European Social Survey (ESS) data. Section 6 provides a summary and discussion on future research directions.

2. Small Area Estimation problem of the Population Mean

2.1 Notation

Given a random sample $s \subset \Omega$ of size n drawn from the target finite population Ω of size N let us denote by $d = 1, \dots, D$ the small areas for which we want to compute the small area estimates. $N - n$ are the non-sampled units and these are denoted by r, hence $s_d = s \cap \Omega_d$ is the sub-sample from the small area d of size n_d, $n = \sum_{d=1}^{D} n_d$, and $s = \cup_d s_d$. r_d denotes the non-sampled units in small area d with $N_d - n_d$ dimension.

The mean for the population Ω of the variable Y for area d is given by the following:

$$\bar{Y}_d = N_d^{-1} \sum_{i=1}^{N_d} y_{di}, \tag{1}$$

with y_{di} denoting the value of variable Y for i^{th} unit from d^{th} area.

2.2 Horvitz-Thompson estimator

A design-unbiased direct estimator to estimate (1) is the Horvitz-Thompson estimator also known as the expansion estimator (Horvitz and Thompson, 1952). This is given by:

$$\hat{\bar{Y}}_d^{Dir} = \frac{\sum_{i \in s_d} y_{di} w_{di}}{\sum_{i \in s_d} w_{di}} \tag{2}$$

with $w_{di} = \pi_{di}^{-1}$ where π_{di} denotes the first-order inclusion probability of i^{th} unit from d^{th} area in s_d. A measure of uncertainty of (2) can be its variance since the estimator is unbiased (Cochran, 1977; Särndal et al., 1992; Rao and Molina, 2015). This is denoted by $Var(\hat{\bar{Y}}_d^{Dir})$ and $\widehat{Var}(\hat{\bar{Y}}_d^{Dir})$ is its estimate in the remainder of the paper.

Due to the unplanned domains problem n_d may be too small in order to be able to compute reliable estimates of (2) using $\hat{\bar{Y}}_d^{Dir}$. Where n_d is equal to zero then $\hat{\bar{Y}}_d^{Dir}$ is in addition not viable. In both circumstances indirect estimation techniques that use auxiliary variables are instead needed (Rao and Molina, 2015).

2.2 Small Area Estimator based on the Iterative Proportional Fitting algorithm

As outlined earlier, one such indirect small area estimator that uses auxiliary variables in this context is the iterative proportional fitting (IPF) algorithm. IPF is widely understood and utilised within the spatial microsimulation family of SAE methods (Ballas et al., 2005; Ballas et al., 2007; Anderson, 2007). and is one of several calibration algorithms based on the minimisation of different distance functions (Deville and Särndal, 1992).

IPF is a reweighting technique used to adjust survey contingency tables of individual characteristics to fit known margins of constraints (usually Census totals) at the small area level. IPF can therefore be considered as a survey weights calibration problem (Deville and Särndal, 1992; Creedy, 2003; Whitworth, 2013; Moretti and Whitworth, 2019b). For each small area, the result of the algorithm is that survey individuals are fractionally reweighted across the selected constraint variables such that they come to represent a synthetic micro-population of each target small area based on its population profile across the constraint variables. For each small area, therefore, tailored weighted are calculated such that their sum equals the small area's population total and the weighted total of the categories across the various auxiliary constraint variables approximates or equals the equivalent actual totals (typically provided by the Census data). An estimate of the target parameter can then be obtained as the weighted combination (usually, but not necessarily, mean or total) of the target outcome variable in the sample based on the final IPF weights.

In order to introduce the IPF algorithm the notation adopted in Kolenikov (2014) is used. Here, w_i denotes the initial weight (this can be the design-weight) for $i \in s_d$. The calibration problem is area-specific such that the IPF algorithm needs to be applied in each area. The IPF algorithm generates calibrated weights denoted by w_i^* for $i \in s_d$ that satisfy the calibration equation given by $\sum_{i \in s_d} w_i^* \boldsymbol{x}_i = \sum_{i \in \Omega_d} \boldsymbol{x}_i = \boldsymbol{X}_d$, where $\boldsymbol{x}_i$ is a vector of auxiliary variables for $d = 1, \dots, D$. In particular, w_i^* minimises a distance function in the case of IPF between $\{w_i^*; i \in s_d\}$ and $\{w_i; i \in s_d\}$. The constrained optimisation problem is given by the following (Deville and Särndal, 1992; Chen and Shen, 2015):

$$\begin{aligned} &min: \sum_{i \in s_d} \left[w_i \ln\left(\frac{w_i}{a_i}\right) - w_i + a_i \right], \\ &such\ that \sum_{i \in s_d} w_i \boldsymbol{x}_i = \sum_{i \in \Omega_d} \boldsymbol{x}_i = \boldsymbol{X}_d. \end{aligned} \tag{3}$$

It is noted that (3) does not have closed-solution, due to the non-linear distance function, and the IPF algorithm is employed to estimate the final calibrated weights.

The steps of the algorithm as follows:

1. Initialize the iteration counter $t \leftarrow 0$ and the weights as $w_i^{0,v} \leftarrow w_i$;
2. Increment the iteration counter $t \leftarrow t + 1$, thus updating the weights as $w_i^{t,0} \leftarrow w_i^{t-1,v}$;
3. Update the weights through the calibration variables $v = 1, \dots, p$:

$$w_i^{t,v} = \begin{cases} w_i^{t,v-1} \dfrac{T(\boldsymbol{X}_v)}{\sum_{l \in s} w_l^{t,v-1} x_{vl}}, & x_{vi} \neq 0 \\ w_i^{t,v-1}, & x_{vi} = 0 \end{cases}.$$

4. If the discrepancies between $\sum_{i \in s} w_i^{t,p} x_v$ (i.e. the sample weighted totals) and $T(\boldsymbol{X}_v)$ are within a priori defined tolerance for all $v = 1, \dots, p$, then declare convergence and the algorithm goes to step 6, otherwise return to step 2;

5. The weights $w_i^{t,p}$ are the final calibrated weights and are denoted by $w_i^{t,p} = w_i^*$.

The variables used for calibration are usually categorical variables in real data applications, therefore we define the following vector:

$$\boldsymbol{x}_i' = \left(\delta_{1i}^{(1)}, \dots, \delta_{F_1 i}^{(1)}, \delta_{1i}^{(2)}, \dots, \delta_{1i}^{(p)}, \dots, \delta_{F_p i}^{(p)}\right),$$

where $l = 1, \dots, p$ denotes the l^{th} control variable and $\delta_{ki}^{(l)} = 1$ if I is in the category k of l^{th} control variable. F_l is the number of categories of the l^{th} control variable. Anderson (2007) suggests that $R = 20$ is ample to satisfy convergence of the algorithm and we follow this cautious advice, though noting that others suggest that fewer iterations may be sufficient (Ballas et al., 2005; Lovelace and Dumont, 2016). In terms of the survey calibration problem framed in formula (3) this means that weights are found that minimise the distance function given the benchmark constraints. This point is also discussed in Lovelace and Dumont (2016). The IPF estimator is defined as follows:

$$\hat{\bar{Y}}_d^{IPF} = \frac{\sum_{i=1}^n w_{di}^* y_i}{\sum_{i=1}^n w_{di}^*}, \quad d = 1, \dots, D, \qquad i = ,1 \dots, n, \tag{4}$$

where w_{di}^* denotes the calibrated survey weight for unit i^{th} from area d^{th}. Note that y_i appears for $i = 1, \dots, n$; this means that $\hat{\bar{Y}}_d^{IPF}$ belongs to the class of synthetic SAE estimators (Rao and Molina, 2015). Of course, considering the case of small n_d, $\hat{\bar{Y}}_d^{IPF}$ is more efficient than $\hat{\bar{Y}}_d^{Dir}$ if the auxiliary variables used in the calibration problem are sufficiently related to the target variable Y (Fuller, 2002 and Moretti and Whitworth, 2019b).

Calibration estimators are known to be model-assisted by which is meant that it is only necessary that the population is reasonably well described by an assumed model in order for that model to be valid for use, this is a property of model-assisted estimators (Särndal et al., 1992; Espuny-Pujol, et al., 2018). Nonetheless, if the model assumptions fail then the gains in efficiency of the IPF estimator compared to a design-based direct estimator may be small. Interestingly, as discussed in Hedlin, et al. (2001) simply because a model-assisted estimator satisfies the property just discussed it may still produce poor estimates. Naturally, the property is not a substitute for a careful model search, particularly in cases where there are outliers or highly variable data (Hedlin, et al., 2001). Biases are expected to grow where model assumptions are not met and this needs to be taken into account when producing estimates (Griffiths, 1996; Berg and Fuller, 2009; Moretti and Whitworth, 2019a; Moretti and Whitworth, 2019b). In this work, we consider a linear model that relates y_{id} to a set of covariates Kott (1990) where the use of this model is considered in model-assisted estimators (Kott, 1990; Moretti and Whitworth, 2019a; Deville, and Särndal, C.E., 1992). In particular, since in small area estimation we aim to consider between area variation the Battese, Harter, Fuller model (Battese et al, 1988) is used and given as follows:

$$\begin{gathered} y_{di} = \boldsymbol{x}_{di}^T \boldsymbol{\beta} + u_d + e_{di}, \qquad i = 1, \dots, N_D, \qquad d = 1, \dots, D, \\ u_d \sim N(0, \sigma_u^2), \; e_{di} \sim N(0, \sigma_e^2 b_{di}^{-1}), \textit{ independent}, \end{gathered} \tag{5}$$

where $\boldsymbol{x}_{di}^T$ denotes a $p \times 1$ column vector of auxiliary variables, $\boldsymbol{\beta}$ denotes the regression coefficients. u_d is the random effect independent to the error term denoted by e_{di}. b_{di} refers to the heteroskedasticity weights, and it is assumed that these are function of the auxiliary

variables, i.e. $b_{di} = b(\boldsymbol{x}_{di})$, if the error are homoskedastic then $b_{di} = 1$ (see also González-Manteiga, et al., 2008).
Like all indirect estimates, IPF estimates is affected by both bias and variance. As such, the mean squared error (MSE) is the appropriate measure of estimator performance as it takes into account both bias and variance. A validated approach to estimate the MSE (4) in a spatial microsimulation context has been proposed in Moretti and Whitworth (2019a) using a parametric bootstrap algorithm. This will be adopted in this paper and is summarised in the Appendix. Evaluations on the role of the model in IPF context have been proposed in the literature, for example in Moretti and Whitworth (2019a). In this latter paper, it is shown that if the errors follow a distribution approximately normally distributed or slightly skewed, then the bias returned by the IPF estimator is negligible (tending to zero). However, heteroskedastic errors might have an important impact on the estimator (Gujarati and Porter, 2009) and this is explored in section 4.

3 Composite Estimators

As noted above, synthetic SAE estimators such as IPF offer the potential to markedly reduce the variance and overall MSE of small area estimates compared to direct survey estimates in contexts where there are small or zero sample sizes across small areas. In turn, composite estimators that take a weighted combination of direct and synthetic (e.g. IPF) estimators offer potential to reduce MSE further still by drawing increased on the direct estimator as small area sample sizes increase and that direct estimator thus becomes more accurate and precise, and vice versa. Given that the focus of this article is on the original development of an optimal composite estimator for spatial microsimulation SAE, this section outlines both that composite estimator and, firstly, the sample-size-dependent composite estimator against which it will be evaluated empirically.

3.1 Sample-size-dependent composite estimator

A sample-size-dependent composite estimator can be defined as follows (Griffiths, 1996; Drew et al, 1982; Moretti and Whitworth, 2019b):

$$\hat{\bar{Y}}_{\delta,d}^{SSD} = \hat{\gamma}_{\delta,d}^{SSD}\hat{\bar{Y}}_d^{Dir} + \left(1 - \hat{\gamma}_{\delta,d}^{SSD}\right)\hat{\bar{Y}}_d^{IPF}, with\ \hat{\gamma}_{\delta,d}^{SSD} = \begin{cases} 1\ if\ \dfrac{n_d}{n} \geq \delta\left(\dfrac{N_d}{N}\right) \\ \left(\dfrac{1}{\delta}\right)\dfrac{n_d/n}{N_d/N}\ if\ \dfrac{n_d}{n} < \delta\left(\dfrac{N_d}{N}\right) \end{cases}, \quad (6)$$

where $\delta \geq 0$ and $0 \leq \hat{\gamma}_{\delta,d}^{SSD} \leq 1$. It can be seen that $\hat{\bar{Y}}_{\delta,d}^{SSD}$ depends on the coefficient δ: when δ increases the effect of borrowing strength from the related small areas increases thus increasing the weighting within the composite estimator that is attached to synthetic IPF estimator and, equivalently, decreasing the weighting attached to the direct estimator (Drew et al, 1982).

3.2 Optimal composite estimator

As pointed in Drew et al. (1982), sample-size-dependent composite estimators were originally developed to deal with small areas for which the sample sizes are large enough such that direct estimators for small areas with sample sizes exceeding the expected sample sizes meet some reliability requirements. Furthermore, sample-size-dependent estimators do not take into account for the between-area heterogeneity, (Rao and Molina, 2015) therefore, if this variable is large the sample-size-dependent estimator might not be efficient.

In contrast, in an optimal composite estimator the weighting of the direct and synthetic components is defined as a function of the MSE of those respective components. As such, given

that minimisation of the MSE is built into its construction optimal composite estimators are in their nature built around the optimisation of performance in the final estimates. An optimal composite estimator of (1) can be defined as follows:

$$\hat{\bar{Y}}_d^{Opt} = \hat{\gamma}_d^{Opt}\hat{\bar{Y}}_d^{Dir} + \left(1 - \hat{\gamma}_d^{Opt}\right)\hat{\bar{Y}}_d^{IPF}, \tag{7}$$

where $0 \leq \hat{\gamma}_d^{Opt} \leq 1$. $\hat{\gamma}_d^{Opt}$ is the optimal weight obtained by minimising the design mean squared error of $\hat{\bar{Y}}_d^{Opt}$. Assuming that the covariance term $E(\hat{\bar{Y}}_d^{Dir} - \bar{Y}_d)(\hat{\bar{Y}}_d^{IPF} - \bar{Y}_d)$ is small relative to $MSE(\hat{\bar{Y}}_d^{IPF})$ we approximate the optimal weight as follows (Schaible, 1978):

$$\hat{\gamma}_d^{Opt} \approx \frac{\widehat{MSE}\left(\hat{\bar{Y}}_d^{IPF}\right)}{\left[\widehat{MSE}\left(\hat{\bar{Y}}_d^{Dir}\right) + \widehat{MSE}\left(\hat{\bar{Y}}_d^{IPF}\right)\right]}. \tag{8}$$

Note that since $\hat{\bar{Y}}_d^{Dir}$ is an unbiased estimator of $\bar{Y}_d$ it is usually assumed that $\widehat{MSE}\left(\hat{\bar{Y}}_d^{Dir}\right) \approx \widehat{Var}\left(\hat{\bar{Y}}_d^{Dir}\right)$ (Rao and Molina, 2015). An estimator of the mean squared error of $\hat{\bar{Y}}_d^{Opt}$, denoted by $MSE(\hat{\bar{Y}}_d^{Opt})$ can be given by the following (Schaible, 1978):

$$\widehat{MSE}\left(\hat{\bar{Y}}_d^{Opt}\right) = \hat{\gamma}_d^{Opt}\widehat{MSE}(\hat{\bar{Y}}_d^{Dir}) = \left(1 - \hat{\gamma}_d^{Opt}\right)\widehat{MSE}\left(\hat{\bar{Y}}_d^{IPF}\right). \tag{9}$$

4 Simulation study

Having set out the small area estimators in Section 3, Section 4 presents results from a simulation study into their relative performance. Specifically, the simulation study evaluates the performance of our new optimal composite spatial microsimulation estimator given by (6) against the sample-size-dependent composite estimator introduced recently into the literature (Moretti and Whitworth, 2019b). Moreover, some initial investigations of an MSE estimator of the optimal composite estimator are presented. This is important in order to provide a good measure of uncertainty of the estimator. The simulation study is a model-based simulation study since model assumptions are relevant to the performance of the IPF and composite small area estimators.

4.1 Simulation study design and population generation

For the simulation study $S = 1{,}000$ populations are generated from the model given in 5.

In order to motivate the use of composite estimators we introduce a mild level of heteroskedasticity in the population. Heteroskedasticity is common in real data applications where for many different reasons users may face to violations of homoscedastic errors such that the variance of e_{di} may not be constant anymore (Gujarati and Porter, 2009). A moderate level of heteroskedasticity is introduced in the simulation by fixing $b_{di}^{\lambda} = 1/x_{di}^{\lambda}$ for $\lambda = 1/2$ following González-Manteiga et al. (2008) in order to produce IPF biased estimate

The simulation is based on an unbalanced population using the following parameters chosen according to Moretti et al. (2018):

- $N = 20{,}000$, $D = 80$, and $130 \leq N_d \leq 420$. N_d, $d = 1, \ldots, D$ is generated from the discrete Uniform distribution ($dUnif$), $N_d \sim dUnif(130, 420)$, with $\sum_{d=1}^{D} N_d = 20{,}000$;
- $\boldsymbol{x}_{di} = (1 \;\; x_{di1} \;\; x_{di2})^T$, with $x_{di1} \sim dUnif(145, 459)$, and $x_{di2} \sim dUnif(55, 345)$,
- $\boldsymbol{\beta} = (17.97 \quad 0.36 \quad -0.03)^T$,

- $\sigma_u^2 = -\frac{\rho}{\rho-1}\sigma_e^2$ and $\sigma_e^2 = 297.71$ with $\rho \in \{0.05, 0.10, 0.20, 0.50, 0.70\}$

where, $\rho = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}$ denotes the intra-class correlation coefficient that partitions the total variance into that which is between-areas and that which is within-areas. The level of the intra-class correlation coefficient is relevant in SAE contexts since the variability of small area estimators depend on this coefficient. The simulation study below is for this reason conducted across multiple levels of intra-class correlation coefficient (see e.g. Moretti et al, 2019; Moretti and Whitworth, 2019a). In the social sciences, the intra-class correlation coefficient does not usually assume such large values. In medical or agricultural applications, however, it can reaches higher values such as these (see. e.g. Koo and Li, 2016; Pleil, et al, 2018). The two auxiliary variables have categories defined as follows in order to identify the constraints used in the IPF algorithm:
$145 \leq x_{1i} \leq 224.20,\ 224.20 < x_{1i} \leq 380.70,\ 380.70 < x_{1i} \leq 459,$
$55 \leq x_{2i} \leq 126.30,\ 126.30 < x_{2i} \leq 272.10,\ 272.10 < x_{2i} \leq 345.$

4.2 Simulation steps

The simulation steps are as follows:

1. Generate the responses y_{dis} according to model (5) for $s = 1, \dots, S$, $(S = 1{,}000)$ with parameters as described above;
2. Draw a stratified random sample (simple random sample without replacement selection in each area d) from each simulated population, with $n_d \sim dUnif(7, 21)$ and $n = \sum_d n_d = 1129$ (the average sampling fraction is $\bar{f}_d = 0.05$);
3. Estimate $\bar{Y}_d$ via the IPF-based estimator given in (4) and the Horvitz-Thompson estimator given in (2). These are denoted by $\hat{\bar{Y}}_{ds}^{IPF}$ and $\hat{\bar{Y}}_{ds}^{Dir}$;
4. Estimate $\bar{Y}_d$ via the composite estimators given by (6) and (7). These are denoted by $\hat{\bar{Y}}_{ds}^{SSD}$ and $\hat{\bar{Y}}_{ds}^{Opt}$ respectively;
5. The results are evaluated using two quality measures where $\hat{\bar{Y}}_{ds}$ denotes any estimate of $\bar{Y}_{ds}$.

The empirical root mean square error (RMSE) offers an overall measure of estimate quality taking into account both variance and bias:

Empirical Root Mean Squared Error

$$RMSE\left(\hat{\bar{Y}}_d\right) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\hat{\bar{Y}}_{ds} - \bar{Y}_{ds}\right)^2}, \tag{10}$$

Relative Bias

$$RB\left(\hat{\bar{Y}}_d\right) = \frac{1}{S}\sum_{s=1}^{S}\frac{\hat{\bar{Y}}_{ds}}{\bar{Y}_{ds}} - 1. \tag{11}$$

Relative bias (RB) is related to the accuracy of an estimator.

4.3 Results

This section presents the main results of the simulation study. The section looks firstly at the results for our proposed original optimal composite estimator for spatial microsimulation approaches to SAE that is the central interest of the article. The section turns next to the results

of the comparison between the optimal composite estimator and the sample-size-dependent composite estimator. The section concludes with preliminary evaluations on an MSE estimator of the optimal composite estimator.

4.3.1 Optimal composite estimator

Table 1 shows the performance of the small area optimal composite estimator $\hat{\bar{Y}}_d^{Opt}$ compared to the IPF estimator $\hat{\bar{Y}}_d^{IPF}$ in terms of relative bias and empirical root mean squared error. The median has been chosen as a robust central tendency measure across the small areas (Chambers et al., 2011; Giusti et al., 2013).

It can be seen that when the intra-class correlation is small the IPF estimator is not biased: the bias across the small areas is negligible for both $\lambda = 0.2$ and $\lambda = 0.5$ cases. However, when the intra-class correlation increases to $\rho = 0.5$ and $\rho = 0.7$ the bias of the IPF estimator increases. The biases are slightly larger when heteroskedasticity increases to $\lambda = 0.5$ as compared to $\lambda = 0.2$.

These scenarios motive the analysis to develop an optimal composite estimator between the direct and IPF estimators in order to explore its potential to reduce this bias seen when the IPF alone is used. Table 1 shows that the optimal composite estimator is indeed able to produce estimates with lower bias than the IPF estimator alone. Naturally, when the intra-class correlation is small and the IPF estimator is relatively unbiased then then performance gap between the IPF estimator and the optimal composite estimator is modest. However, as the intra-class correlation increases and the bias of the IPF estimator becomes larger this is not the case with the optimal composite estimator which continues to produce more precise (variance) and accurate (bias) estimates by giving more weight to the unbiased direct estimates at these points. These results can be seen particularly clearly when $\rho = 0.5$ and $\rho = 0.7$.

<Table 1 about here>

4.3.2 Comparisons between optimal composite estimator and sample-size-dependent estimator

Having evaluated the performance of our optimal composite estimator against the IPF estimator alone, Table 2 turns next to the comparison of the optimal composite estimator and the sample-size-dependent composite estimator recently introduced to the spatial microsimulation small area estimation literature (Moretti and Whitworth, 2019b). The first two rows of Table 2 show the relative bias and the empirical root mean squared error of the optimal composite estimator ($\hat{\bar{Y}}_d^{Opt}$). The remaining rows of Table 2 show the relative bias and empirical root mean squared error of the sample-size-dependent estimator ($\hat{\bar{Y}}_{d,\delta=\delta^*}^{SSD}$) at different levels of weightings to combine its direct and synthetic components, $\delta^* = \left\{0.2, 0.5, \frac{2}{3}, 1, 1.2, 2, 2.5, 3, 3.5, 4, 4.5, 5\right\}$.

<Table 2 about here>

Looking across Table 2 it can be seen that the optimal composite estimator performs better than the sample-size-dependent estimator in all scenarios across these key performance metrics. As one would hope from an optimal composite estimator it is indeed optimal in the sense that there is no scenario in which the combination of estimated root mean square error and relative bias is superior in the sample-size-dependent estimator than the combination obtained from the optimal composite estimator. The smallest estimated root square error that can be obtain from

the sample-size-dependent composite estimator estimates is when $\delta = 5$. However, the relative bias returned at this point by that sample-size-dependent composite estimator is larger than that of the optimal composite estimator. This is due to the fact that when δ increases more weight is attached to the biased IPF estimator within that sample-size-dependent estimator.

4.3.3 On the MSE of $\hat{\bar{Y}}_d^{Opt}$

Here we present the results of a preliminary attempt to provide a measure of uncertainty of the optimal composite estimator.

Table 3 shows the evaluation of the MSE estimator of the optimal composite estimator, denoted by $\widehat{RMSE}\left(\hat{\bar{Y}}_d^{Opt}\right)$ estimated via (8). The results are presented in terms of root mean squared error. The first line of Table 3 is about the empirical root mean squared error (the true), the second line contains its estimate and the third line shows the bias of $\widehat{RMSE}\left(\hat{\bar{Y}}_d^{Opt}\right)$.

<Table 3 about here>

It can be seen that, estimator $\widehat{MSE}\left(\hat{\bar{Y}}_d^{Opt}\right)$, given in (8), returns good estimates of the empirical mean squared error of $\hat{\bar{Y}}_d^{Opt}$, the bias across the small areas is small. However, in some cases the mean squared error is slightly overestimated, i.e. for $\lambda = 0.2$ with $\rho = \{0.05, 0.5, 0.7\}$ and $\lambda = 0.5$ with $\rho = \{0.05, 0.1, 0.7\}$, in the other cases it is slightly underestimated. More investigations on this will be object of future work particularly considering resampling techniques also.

5 Application

Section 4 has evaluated findings of a simulation study across different small area estimators and found our proposed composite estimator for spatial microsimulation SAE to deliver the best levels of performance in terms both of the minimisation of bias and variance around the final estimates as well as the solid estimation of the mean squared error. To aid understanding and dissemination of the optimal composite estimator in such contexts Section 5 presents an application based on real data using Italian data from the European Social Survey (ESS).

5.1 The data

Data from Italy in the 8th round of the ESS are used in this application. ESS samples are representative of all persons aged 15 and over resident within private households in each country, regardless of their nationality, citizenship or language. The sampling design for Italy is a two domain design. The first sampling domain consists of the nine biggest municipalities within Italy. For these municipalities a one-stage sampling design is used where a total of 770 individuals are sampled using simple random stratified sampling where the sample size allocation is proportional to the target population in the strata. The second sampling domains consist of all other municipalities and here a two-stage sampling design is used. In the first stage 163 municipalities are selected as Primary Sampling Units (PSUs) by stratified sample based on the crossing of NUTS-1 geographies and demographic profile of the surveyed population (4 classes). The allocation of the PSU sample to the strata is proportional to target population within the strata. In the second stage 29 individuals are selected from each sampled municipality using a simple random sampling. The survey documentation states that reliable statistical inference is not advised at regional (NUTS-2) level (European Social Survey, 2017) due to small survey sample sizes at these smaller area geographies. Hence, a small area estimation problem exists for users wishing spatial granularity in ESS indicators.

5.2 A optimal composite estimate of trust

This application of the optimal composite estimator focuses on the small area estimation of political trust in Italy. As discussed in André (2014) political trust is an important issue in contemporary representative democracy not only in and of itself but also given that it can support wider social trust, foster associational life and play an important role in the efficient implementation of policies.

André (2014) discusses the ways in which political trust can be considered a multidimensional phenomenon such that multiple dimensions need to be considered in any measurement of the latent concept. In terms of suitable indicators André (2014) suggests the following questions of the ESS round 8 data: "Using this card, please tell me on a score of 0-10 how much you personally trust each of the institutions I read out. 0 means you do not trust an institution at all, and 10 means you have complete trust":

- parliament
- the legal system
- the police
- politicians
- political parties
- the European Parliament
- United Nations.

These responses make up the observed target variables for our application. For the purposes of this application these indicators are brought together using principal components analysis (PCA) into a set of non-correlated linear combinations. It is noted that other multivariate statistical analysis techniques may alternatively be used to handle this multi-dimensionality but this is both beyond and not the purpose in this application. Only one of the components produced from the PCA carry an eigenvalue above 1 and this first component alone explains 64.3% of the total variance across these indicators. This first component is therefore taken as the target outcome measure of the multi-dimensional latent concept of political trust for the purposes of this application.

The target parameter of this application is the mean of the trust indicator and the aim is to produce reliable estimate at the regional (NUTS-2) level in Italy that the ESS survey document advises cannot reliably be delivered directly. As such, although these twenty Italian regions are not 'small areas geographically they retain the core SAE problem of being unplanned domains in ESS data given that the survey sampling strategy was designed with reliable statistical inference viable only at higher level geographies and not possible to this spatial scale.

In this application the mean of the political trust score is estimated to regional level using our proposed optimal composite estimator (6) combining the direct estimator (2) and the IPF estimator (4). The auxiliary variables used as constraints to provide IPF estimates are the following: marital status, age, citizen of Italy, level education (EISCED scale), employment status (working or not), currently enrolled in a course (school or university). It will be remembered that the variance of the direct estimator and the MSE of the IPF estimator are needed in order to provide the optimal small area estimates since seek to balance for the large variability but zero bias of the direct estimates and the bias but smaller variability from the IPF estimates. These are referred to as measures of uncertainty here. The MSE must be considered for the IPF estimator since this takes into account both variance and bias ($MSE = VAR + BIAS^2$) (see Rao and Molina, 2015). The variance of the direct estimator is estimated according to Särndal et al. (1992), the mean squared error (MSE) of the small area IPF estimator is

estimated via parametric bootstrap according to Moretti and Whitworth (2019a) and the MSE of the optimal composite estimator is calculated according to formula (7).

Figure 1 compares the performance of three estimators in producing these regional estimates of political trust: the direct estimator, the IPF estimator alone, and our proposed optimal composite estimator combined the two with optimal weights. It can be seen from Figure 1 that the optimal composite and IPF estimators both provide better performance than the direct estimator for regions with smaller survey sample sizes (towards the left of Figure 1). As regions survey sample sizes increase the performance of the direct estimator naturally improves as its variance decreases. The performance of the IPF estimator remains roughly stable, however, such that the direct estimator begins to outperform the IPF estimator at larger survey sample sizes. For the optimal composite estimator, in contrast, this always performs better than both the direct estimator and the IPF estimator across all region survey same sizes. These evaluations are necessary in order to evaluate the quality of the small area estimates as from guideline from official statistical institutes, in fact, MSE and variance estimates are considered as measures of statistical quality; we refer to Statistics Canada (2009) for details on this topic in official statistics. Furthermore, it is crucial to remind that as from ESS guidelines, reliable statistical inference is not advised at regional (NUTS-2) level for Italy (European Social Survey, 2017).

<Figure 1 about here>

In order to evaluate the quality of the small area estimates these can be compared to the direct estimates that are design-unbiased (but with large variance). Thus, according to the small area estimation literature e.g. Moretti and Whitworth (2019b) and Brown et al. (2001) we can estimate simple bivariate linear regression models where the direct estimate denotes the dependent variable and the IPF or optimal estimate denote the independent variable. Here the estimates of the regression coefficients are the following $\hat{\beta}_0 = 0.44, \hat{\beta}_1 = 0.30$ and $\hat{\beta}_0 = 0.01, \hat{\beta}_1 = 1.02$ for the IPF and optimal composite estimates, respectively. These demonstrate that the optimal composite estimates both display relatively little bias in an absolute sense and display markedly less bias than the IPF estimates. These results are in line with the simulation study findings.

Figure 2 shows the maps of the mean regional estimates of the political trust indicator for both the IPF estimator (left) and the optimal composite estimator (right).

<Figure 2 about here>

Figure 2 shows noticeable variability in the regional estimates of political trust between different Italian regions. The highest levels of political trust are estimated across the northern regions including Lombardia, Trentino Alto Adige and Friuli Venezia Giulia, though with some northern exceptions such as Valle d'Aosta, Liguria and Veneto. In the centre of Italy high levels of trust can generally be seen, particularly in Tuscany and Umbria. Across southern regions high levels of trust are estimated in Molise and Basilicata and medium levels in Sicilia and Calabria. As noted in Fazio et al. (2017), these maps highlights that the geographical distribution of political trust across Italy is more complex than the simplistic the North–South divide that is often used to describe Italy.

Although the estimates are, as one would expect, broadly comparable between the IPF and optimal composite estimators there remain several points of important difference such as in Lazio, Puglia and Sicilia. This highlights the importance of the choice of estimator not only to

its performance in terms of minimisation of bias and variance as has been the focus above but also in terms of its implications for the central point estimates themselves.
With this application, we can show that we are able to provide reliable regional estimates of the phenomenon of interest, with a smaller uncertainty than the direct and IPF estimates. IPF estimates, due to their possible small area biases, are not always reliable. Thus, by constructing optimal small area estimates, it is possible to produce more reliable small area estimates for every region (see Figure 1).

6 Conclusion

Spatial microsimulation approaches to small area estimation are widely practised by research and policy analysts in order to estimate new indicators at finer spatial resolution in order to enhance our spatial understanding of societies and policy interventions. As with any synthetic estimator, however, bias is introduced. Whilst regression-based SAE approaches have for some time made use of the potential for composite estimators to help optimise the balance between bias and variance minimisation this is largely neglected within spatial microsimulation SAE approaches despite their widespread use and popularity.

The potential of the sample-size-dependent composite estimator to improve the quality of spatial microsimulation approaches to small area estimation has recently been proposed and evaluated positively in the literature (Moretti and Whitworth, 2019b). Whilst an important contribution, such sample-size-dependent composite estimators suffer from neglecting between-area variation. Therefore, if there is a large heterogeneity between small areas, sample-size-dependent estimators might not be much more efficient compared to direct design-based estimators. The present article builds on that recent advance by pushing further the quality of such approaches through its original development and empirical evaluation of optimal composite estimators within a spatial microsimulation SAE framework for the first time in the SAE literature under microsimulation approaches. Unlike sample-size-dependent composite estimators, optimal composite estimators are so named because the key weighting between the direct and synthetic parts of the estimator is derived explicitly from the minimisation of their respective variance and bias as is desired in order to maximise the performance of the estimator overall.

After having set out the notation of our proposed composite estimator in this context, the article's empirical findings show that the optimal composite estimator produced superior performance to either the direct, IPF or sample-size-dependent estimator across all levels of intra-class correlation coefficient, heteroskedasticity and small are survey sample size. The bias of the composite estimator is close to zero in all the scenarios examined and that a mean squared error estimate based on an expression in Schaible (1978) successfully approximates the actual empirical mean squared error. Taken together the results offer strong evidence to suggest that widely used spatial microsimulation approaches to SAE should give strong consideration to instead utilising that spatial microsimulation approach as the synthetic element alongside a direct estimator within a larger optimal composite estimator. Of course, if the sample size in the small area is zero then direct estimation techniques cannot be applied and researchers need to rely on the synthetic IPF estimator. This is common to all the small area estimation techniques available in the literature. For more details on this we refer to Rao and Molina (2015) where this point is discussed in the regression-based small area estimation context.

Future research could usefully focus on new methods (e.g. via resampling techniques) to estimate the mean squared error of the optimal composite estimator. In addition, further work could explore the performance of the proposed optimal composite estimator under a wider

range of different scenarios in terms of dimensions such as data size and normality, failures in model assumptions, spatial variability of the auxiliary variables and non-linear outcome variables. Other interesting extensions of this work would be applications of the optimal composite estimator to other spatial microsimulation techniques as well as alongside regression-based techniques in order to offer comparative methodological evidence for SAE practitioners around the performance implications of different synthetic specifications within optimal composite SAE estimators.

References

Anderson, B. (2007). Creating small area income estimates for England: spatial microsimulation modelling, a report to the Department of Communities and Local Government. London: Department of Communities and Local Government (available at https://webarchive.nationalarchives.gov.uk/20120919223716/http://www.communities.gov.uk/documents/communities/pdf/325286.pdf).

André, S. (2014). Does trust mean the same for migrants and natives? Testing measurement models of political trust with multi-group confirmatory factor analysis. Social Indicators Research, 115(3), 963-982.

Ballas, D., Clarke, G., Dorling D and Rossiter, D (2007). Using SimBritain to Model the Geographical Impact of National Government Policies, Geographical Analysis 39(1): 44-77.

Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B. and Rossiter, D (2005). SimBritain: A Spatial Microsimulation Approach to Population Dynamics. Population, Space and Place 11: 13-34.

Bartholomew, D. J., Steele, F., Galbraith, J., and Moustaki, I. (2008). Analysis of multivariate social science data. Chapman and Hall/CRC.

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An Error-Components model for Prediction of County crop areas using Survey and Satellite data. Journal of the American Statistical Association 83(401): 28-36.

Berg, E., and Fuller, W. A. (2009). A SPREE Small Area Procedure for Estimating Population Counts. Proceedings of the Survey Methods Section, SSC Annual Meeting.

Brown, G., R. Chambers, P. Heady, and D. Heasman. 2001. Evaluation of small area estimation methods – an application to the unemployment estimates from the UK LFS. In Statistics Canada Symposium Ottawa, October 2001.

Buil-Gil, D., Moretti, A., Shlomo, N., and Medina, J. (2019). Worry about crime in Europe. A model-based small area estimation from the European Social Survey. European Journal of Criminology.

Chen, H. and Shen, R. (2015). Variance estimation for survey-weighted data using bootstrap resampling methods: 2013 methods-of-payment survey questionnaire. Technical Reports 104, Bank of Canada. Retrieved from https://ideas.repec.org/s/bca/bocatr.html.

Cochran (1977) Sampling Techniques. New York: John Wiley & Sons, Inc.

Creedy, J. (2003). Survey reweighting for tax microsimulation modelling. New Zealand Treasury, Working paper 03/17

Datta, G. S., Day, B. and Basawa, I. (1999) Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. Journal of Statistical Planning and Inference, 75, 269-279.

Deville, J.C. and Särndal, C.E. (1992), Calibration estimators in survey sampling, Journal of the American Statistical Association, 87(418).

Drew, J. D., Singh, M. P., and Choudhry, G. H. (1982). Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey 8: 17-48.

Espuny-Pujol, F., Morrissey, K., and Williamson, P. (2018). A global optimisation approach to range-restricted survey calibration, Statistics and Computing, 28, 427-439.

European Social Survey (2017). ESS Round 8 (2016/2017) Technical Report. London: ESS ERIC.

Fazio, G., Giambona, F., Vassallo, E., & Vassiliadis, E. (2017). A Measure of Trust: The Italian Regional Divide in a Latent Class Approach. Social Indicators Research, 1-34.

Fuller, W.A. (2002). Regression estimation for survey samples, Survey Methodology, 28(1), 5–23.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. and Santamaría, L. (2008). Bootstrap mean squared error of a small-area EBLUP, Journal of Statistical Computation and Simulation, 78, 443-462.
Griffiths, R. (1996). Current Population Survey Small Area Estimation for Congressional Districts. Proceedings of the Section on Survey Research Methods, Washington, DC: American Statistical Association: 314-319.
Hedlin, D., Falvey, H., Chambers, R., and Kokic, P. (2001). Does the model matter for GREG estimation? A business survey example. Journal of Official Statistics, 17(4), 527-544.
Horvitz D.G. and Thompson, D. J. (1952). A Generalization of Sampling without Replacement from Finite Universe, Journal of the American Statistical Association, 47, 663-685.
Ipsos MORI (2015). Multi-level modelling and small area estimation work for the National Survey for Wales. London: Ipsos MORI.
Koch, G.G. (1982). Intraclass correlation coefficient. In Kotz, S. and Johnson, N. L. Encyclopedia of Statistical Sciences. 4. New York: John Wiley & Sons. 213–21.
Kott, P.S. 2009. Calibration Weighting: Combining Probability Samples and Linear Prediction Models in Pfeffermann, D. and Rao, C.R. 2009) Handbook of Statistics 29B Sample Surveys: Inference and Analysis.
Kolenikov, S. (2014) Calibrating survey data using iterative proportional fitting (raking), The Stata Journal, 14(1), 22-59.
Koo, T. K., and Li, M. Y. (2016). A Guideline of Selecting and ReportingIntraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine, 15, 155-163.
Lovelace, R. and Dumont, M. (2016). Spatial Microsimulation with R. Chapman and Hall/CRC.
Maiti, H. (2005). Bayesian Aspects of Small Area Estimation. Handbook of Statistics, 25, 965-982.
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. Computational Statistics and Data analysis 56:2889-2902.
Moretti, A. and Whitworth, A. (2019, forthcoming). Estimating the Uncertainty of a Small Area Estimator based on a Microsimulation Approach. Sociological Methods & Research.
Moretti, A. and Whitworth, A. (2019). Evaluations of small area composite estimators under a microsimulation approach. Communications in Statistics – Simulation and Computation.
Moretti, A., Shlomo, N., and Sakshaug, J.W. (2018) Parametric Bootstrap Mean Squared Error of a Small Area Multivariate EBLUP. Communications in Statistics – Simulation and Computation.
Moretti, A., Shlomo, N., and Sakshaug, J.W. (2019) Multivariate Small Area Estimation of Multidimensional Latent Economic Wellbeing Indicators. To appear in International Statistical Review.
Pleil, J. D., Geer Wallace, A., Stiegel, M. and Funk, W. (2018). Human biomarker interpretation: the importance of intra-class correlation coefficients (ICC) and their calculations based on mixed models, ANOVA, and variance estimates. Journal of Toxicology and Environmental Health, Part B, 21(3), 161-180.
Rahman, A. and Harding, A. (2017). Small Area Estimation and Microsimulation Modeling CRC Press Taylor & Frencis Group.
Rao, J. N. K. and Molina, I. (2015). Small area estimation. New York: Wiley.
Särndal, C.E., Swensson, B., and Wretman, J. (1992). Model assisted survey sampling. Springer.
Schaible, W.L. (1978) Choosing Weights for Composite Estimators for Small Area Statistics (Available at http://www.asasrms.org/Proceedings/papers/1978_159.pdf).
Singh, A and Mohl, C (1996). Understanding calibration estimators in survey sampling, Survey Methodology 22: 107-115.
Statistics Canada (2009) Statistics Canada Quality Guidelines. Catalogue no. 12-539-X.

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

Williamson, P., Birkin, M and Rees, P (1998). The estimation of population microdata using data from small area statistics and samples of anonymised records. Environment and Planning A 30: 785-816.

Wu, C. and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data, Journal of American Statistical Association, 96, 185-193.

Appendix: Mean Squared Error of the Small Area IPF Estimator by Moretti and Whitworth (2019a)

The algorithm steps for the bootstrap MSE for IPF are listed below for $b = 1, \dots, B$ bootstrap replications where the symbol * is used to denote the bootstrap quantities and for $d = 1, \dots, D$ small areas:

1. Fit model (5) to the observed sample data, denoted by s, and estimate the model parameters. The estimates are denoted by $\widehat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$, $\hat{\sigma}_e^2$;
2. Generate the bootstrap area effects $u_d^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_u^2)$;
3. Generate the bootstrap residual error term $e_{di}^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_e^2)$, independently of $u_d^{*(b)}$, for every unit i in the sample in area d, for the sample units, $i \in s_d$;
4. Calculate the true population means for each small area of the bootstrap population as follows:

$$\bar{Y}_d^{*(b)} = \bar{\boldsymbol{x}}_{d,pop}^T \widehat{\boldsymbol{\beta}} + \boldsymbol{u}_d^{*(b)},$$

where $\bar{\boldsymbol{x}}_{d,pop}$ denotes the means of the known population auxiliary variables for each area d. These may be taken, for instance, from the census or administrative data.

5. Generate the bootstrap data as follows, $i \in s_d$:

$$y_{di}^{*(b)} = \boldsymbol{x}_{di}^T \widehat{\boldsymbol{\beta}} + \boldsymbol{u}_d^{*(b)} + \boldsymbol{e}_{di}^{*(b)},$$

6. Compute the IPF estimator on $y_{di}^{*(b)}$ and obtain the IPF estimates on the bootstrap data $\hat{\bar{Y}}_d^{IPF*(b)}$;
7. Repeat steps 2. through 6. for $b = 1, \dots, B$ for each area $d = 1, \dots, D$.

An estimator of $MSE\left(\hat{\bar{Y}}_d^{IPF}\right)$ is given by the following Monte Carlo approximation:

$$\widehat{MSE}_{boot}\left(\hat{\bar{Y}}_d^{IPF}\right) = B^{-1} \sum_{b=1}^{B} \left(\hat{\bar{Y}}_d^{IPF*(b)} - \bar{Y}_d^{*(b)}\right)^2.$$

List of Tables

λ	0.2					0.5				
ρ	**0.05**	**0.1**	**0.2**	**0.5**	**0.7**	**0.05**	**0.1**	**0.2**	**0.5**	**0.7**
$RB\left(\hat{\bar{Y}}_d^{Opt}\right)\%$	0.01	0.08	0.08	0.79	0.32	-0.06	0.30	0.10	0.10	0.41
$RB\left(\hat{\bar{Y}}_d^{IPF}\right)\%$	0.02	0.26	0.41	2.10	5.40	-0.10	0.33	1.10	2.24	6.00
$ERMSE\left(\hat{\bar{Y}}_d^{Opt}\right)$	4.81	6.44	8.36	15.96	21.91	8.41	9.53	10.68	18.53	24.77
$ERMSE\left(\hat{\bar{Y}}_d^{IPF}\right)$	4.81	6.49	8.47	18.54	22.30	7.79	8.67	10.35	18.52	26.95

Table 1 Evaluations of the optimal composite estimator $\hat{\bar{Y}}_d^{Opt}$ compared to the IPF estimator $\hat{\bar{Y}}_d^{IPF}$

λ	0.2					0.5				
ρ	**0.05**	**0.1**	**0.2**	**0.5**	**0.7**	**0.05**	**0.1**	**0.2**	**0.5**	**0.7**
$RB\left(\hat{\bar{Y}}_d^{Opt}\right)$	0.01	0.08	0.08	0.79	0.32	-0.06	0.30	0.10	0.10	0.41
$ERMSE\left(\hat{\bar{Y}}_d^{Opt}\right)$	4.81	6.44	8.36	15.96	21.91	8.41	9.53	10.68	18.53	24.77
$RB\left(\hat{\bar{Y}}_{d\,\delta=0.2}^{SSD}\right)$	0.05	0.14	0.10	0.11	0.16	-0.20	0.18	0.11	-0.05	-0.05
$RB\left(\hat{\bar{Y}}_{d\,\delta=0.5}^{SSD}\right)$	0.05	0.18	0.13	0.22	0.30	-0.20	0.21	0.23	0.08	0.41
$RB\left(\hat{\bar{Y}}_{d\,\delta=2/3}^{SSD}\right)$	0.07	0.19	0.31	0.39	0.46	-0.04	0.22	0.32	0.43	0.75
$RB\left(\hat{\bar{Y}}_{d\,\delta=1}^{SSD}\right)$	0.07	0.19	0.23	0.39	0.66	-0.08	0.25	0.29	0.46	0.87
$RB\left(\hat{\bar{Y}}_{d\,\delta=1.2}^{SSD}\right)$	0.09	0.23	0.18	0.47	0.93	-0.07	0.21	0.34	0.64	1.05
$RB\left(\hat{\bar{Y}}_{d\,\delta=2}^{SSD}\right)$	0.10	0.27	0.27	1.01	2.70	-0.07	0.28	0.39	0.89	2.56
$RB\left(\hat{\bar{Y}}_{d\,\delta=2.5}^{SSD}\right)$	0.12	0.25	0.32	1.21	3.26	-0.13	0.31	0.51	1.15	3.26
$RB\left(\hat{\bar{Y}}_{d\,\delta=3}^{SSD}\right)$	0.14	0.27	0.35	1.34	3.68	-0.12	0.34	0.55	1.31	3.76
$RB\left(\hat{\bar{Y}}_{d\,\delta=3.5}^{SSD}\right)$	0.16	0.28	0.33	1.45	3.92	-0.12	0.34	0.55	1.46	3.98
$RB\left(\hat{\bar{Y}}_{d\,\delta=4}^{SSD}\right)$	0.18	0.27	0.33	1.52	4.12	-0.11	0.33	0.59	1.56	4.23
$RB\left(\hat{\bar{Y}}_{d\,\delta=4.5}^{SSD}\right)$	0.18	0.27	0.32	1.59	4.28	-0.12	0.34	0.59	1.60	4.42
$RB\left(\hat{\bar{Y}}_{d\,\delta=5}^{SSD}\right)$	0.20	0.26	0.35	1.67	4.41	-0.12	0.34	0.59	1.67	4.59
$ERMSE\left(\hat{\bar{Y}}_{d\,\delta=0.2}^{SSD}\right)$	19.07	18.84	19.03	18.54	18.54	33.96	33.96	32.17	33.27	32.66
$ERMSE\left(\hat{\bar{Y}}_{d\,\delta=0.5}^{SSD}\right)$	17.48	17.47	17.80	17.33	17.96	30.99	31.18	30.18	30.97	30.92
$ERMSE\left(\hat{\bar{Y}}_{d\,\delta=2/3}^{SSD}\right)$	13.59	13.73	13.63	14.13	18.51	23.24	24.37	24.16	24.28	23.93
$ERMSE\left(\hat{\bar{Y}}_{d\,\delta=1}^{SSD}\right)$	14.32	14.07	14.18	14.64	16.04	25.03	25.08	24.84	25.24	25.30
$ERMSE\left(\hat{\bar{Y}}_{d\,\delta=1.2}^{SSD}\right)$	13.10	13.17	13.38	13.72	15.64	22.59	23.37	22.93	23.62	23.80
$ERMSE\left(\hat{\bar{Y}}_{d\,\delta=2}^{SSD}\right)$	8.82	9.10	9.38	11.97	14.55	15.46	15.65	15.38	17.41	20.01
$ERMSE\left(\hat{\bar{Y}}_{d\,\delta=2.5}^{SSD}\right)$	7.26	7.51	8.08	11.55	14.86	12.75	13.08	13.13	15.51	18.96
$ERMSE\left(\hat{\bar{Y}}_{d\,\delta=3}^{SSD}\right)$	6.33	6.79	7.37	11.32	15.44	11.10	11.30	11.76	14.79	18.56
$ERMSE\left(\hat{\bar{Y}}_{d\,\delta=3.5}^{SSD}\right)$	5.70	6.21	7.01	11.27	16.02	9.98	10.15	10.71	14.39	18.51
$ERMSE\left(\hat{\bar{Y}}_{d\,\delta=4}^{SSD}\right)$	5.25	5.86	6.80	11.34	16.52	9.11	9.38	10.02	14.08	18.40
$ERMSE\left(\hat{\bar{Y}}_{d\,\delta=4.5}^{SSD}\right)$	4.95	5.62	6.64	11.54	16.93	8.46	8.80	9.56	13.90	18.50
$ERMSE\left(\hat{\bar{Y}}_{d\,\delta=5}^{SSD}\right)$	4.69	5.45	6.57	11.67	17.25	7.97	8.40	9.17	13.76	18.70

Table 2 Performance comparisons of the optimal and sample-size-dependent composite estimators

λ	0.2					0.5				
ρ	**0.05**	**0.1**	**0.2**	**0.5**	**0.7**	**0.05**	**0.1**	**0.2**	**0.5**	**0.7**
$ERMSE\left(\widehat{\bar{Y}}_d^{Opt}\right)$	4.81	6.44	8.36	15.96	21.91	8.41	9.53	10.68	18.53	24.77
$\widehat{RMSE}\left(\widehat{\bar{Y}}_d^{Opt}\right)$	5.12	6.31	8.16	16.51	23.85	8.39	9.86	10.28	17.00	23.75
$RB\left(\widehat{RMSE}\left(\widehat{\bar{Y}}_d^{Opt}\right)\right)$	8.05	-3.93	-4.96	6.15	16.89	1.44	7.87	-8.27	-7.26	-9.69

Table 3 Evaluation of the MSE estimator of the optimal composite estimator