Manchester Metropolitan University

e-space

Manchester Metropolitan University's Research Repository

Please cite the Published Version

Moretti, A and Whitworth, A (2020) Evaluations of small area composite estimators based on the iterative proportional fitting algorithm. Communications in Statistics: Simulation and Computation, 49 (12). pp. 3093-3110. ISSN 0361-0918

DOI: https://doi.org/10.1080/03610918.2018.1535067

Publisher: Taylor & Francis

Version: Accepted Version

Downloaded from: https://e-space.mmu.ac.uk/625176/

Additional Information: This is an Author Accepted Manuscript of an article published in Communications in Statistics: Simulation and Computation by Taylor & Francis.

Enquiries:
If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines)

Evaluations of Small Area Composite Estimators based on the Iterative Proportional Fitting Algorithm

Angelo Moretti[1] and Adam Whitworth[2]

Department of Geography, University of Sheffield, Sheffield, S10 2TN, United Kingdom

1. a.moretti@sheffield.ac.uk and a.moretti2@outlook.com (Corresponding author)
2. adam.whitworth@sheffield.ac.uk

Evaluations of Small Area Composite Estimators based on the Iterative Proportional Fitting Algorithm

Manuscript

Abstract

This article deals with the use of sample size dependent composite estimators in spatial microsimulation approaches for small area estimation. This approach has been applied to regression-based small area estimation approaches but never to our knowledge to spatial microsimulation approaches. In this paper, we extend the iterative proportional fitting (IPF) spatial microsimulation technique to small area composite estimators. Using a simulation study, we show both the impact of sample size and the gains from composite estimation to the mean squared error of IPF-based composite estimators. The target variable used is a binary variable reporting good health or bad health.

Keywords: Small area estimation; spatial microsimulation; IPF; composite estimator; synthetic estimator.

1. Introduction

A wide range of social phenomena such as fear of crime, wellbeing, social exclusion or even income in many contexts are spatially heterogeneous, of interest to policy makers and analysts, yet unavailable at small area level from either census or administrative sources that might offer robust data at that small area scale. Whilst such data are often available from large-scale surveys, providing an invaluable source of rich understanding at larger spatial unit such as the country or regional level, such surveys are not designed to be representative at small area level, and the data collection costs of doing so would be prohibitively high. The problem of unplanned domains thus arises whereby one faces the limitation or small or zero survey sample size at the small area level (Rao and Molina, 2015). In this case the use of the direct design-based estimators using only the sample survey units, such the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), provides a large variability in the estimates or, in zero sample size small areas, no estimation possibility (Rao and Molina, 2015).

This is a significant limitation of direct estimators from large-scale surveys down to small area level both analytically and for policy makers. Small area understanding continues to be – indeed, is increasingly – wanted and demanded by different policy and analysts for at least three key purposes: in a *static* sense to simply understand the differential nature of small areas across a larger territory (typically a nation); in a *dynamic* sense in terms of how the nature of small areas across such characteristics change over time; and in a *policy* sense to assess how small areas respond to policy interventions, whether directly spatially targeted or indirectly spatially affecting (Pratesi and Salvati, 2016).

In this context, small area estimation (SAE) techniques that make use of area auxiliary data from the Census or other administrative data have shown that they offer the ability to overcome the estimation problem by borrowing strength from related areas and auxiliary variables to produce reliable estimates at small area level from large-scale sample survey data (Rao and Molina, 2015). Whilst terminology varies, previous methodological reviews group SAE methodologies into two broad approaches – spatial microsimulation and regression-based statistical modelling – and each with methodological variants within it (Whitworth, 2013; Marshall, 2010). Whilst each methodological approach varies in its application, common across all approaches is the desire to use the relationships seen in the sample survey between explanatory variables and the target outcome variable in order to estimate a target population parameters (such as means, totals, etc.) of that outcome variable.

Associated primarily with social and survey statisticians, regression-based SAE approaches extend the simple within-sample predictive approach to the small area level out-of-sample situation. We refer to Rao and Molina (2015) for a helpful review. A range of model specifications have been adopted including ecological (Ipsos MORI, 2015), mixed-effect (Battese et al, 1988), multivariate mixed-effect (Datta, et al 1999) and M-Quantile (Marchetti et al, 2012) models. Associated mainly with the discipline of quantitative geography, three main spatial microsimulation approaches in contrast involve either the optimal reweighting – iterative proportional fitting (Ballas et al, 2005) and generalized regression (Singh and Mohl, 1996) – or the optimal selection – combinatorial optimisation (Williamson, et al, 1998) – of sample survey cases to fit to the small area profile.

Although small area estimation has been demonstrated to provide acceptably precise estimates from large-scale sample survey data down to the at the small area level, the bias of small area estimates should be carefully taken into account. The threat of bias diminishes however as the small area sample size increases, offering the potential for well informed composite estimators to enhance the small area estimation by combining the indirect small area estimator with the direct sample survey estimator with gradually shifting weights towards the direct estimate as the sample size in the target small area increases. The efficiency of a composite estimator here is measured in the standard way by the reduction in the mean square error (MSE), hence taking into account both bias and variance in the estimation, compared to the variance of a direct estimator when can be calculated simultaneously. Although the opportunities afforded by composite estimation are utilised within regression-based approaches to small area estimation (Rao and Molina, 2015) they have ever been explored within any of the spatial microsimulation approaches, despite those approaches continuing to be widely used utilised to generate small area estimates for both practitioner and academic users. This article rectifies that gap by exploring the potential viability of, and gains from, well informed composite estimation within the popular iterative proportional fitting (IPF) spatial microsimulation approach to small area estimation, also referred to under the name of Structure PREserving Estimation (SPREE) (Zhang and Giusti, 2016) or raking (Deville et al, 1993) within the small area estimation literature (see Purcell and Kish, 1980 for theoretical aspects).

IPF is a deterministic reweighting technique used to adjust contingency tables to fit known margins of constraints at the small areal level – the small area totals of the identified set of explanatory variables typically derived from census data. The result is that survey individuals are reweighted across the selected constraint variables such that they come to represent a synthetic micro-population that is fitted to the characteristics of the small area as seen across the constraint totals. By doing so, the IPF algorithm delivers a set of reweighted survey cases where the number of weighted individuals in total and in the specified categories of the constraint fits to the profile of each target small area, with each small area naturally having its own tailored set of reweights specific to its particular small area profile across the constraints. IPF can therefore be considered as a survey weights calibration problem (see Creedy, 2003).

The statistical properties of IPF are known and studied in the literature (Ballas et al., 2005; Ballas et al., 2007; Anderson, 2007). Interestingly, Agresti (2002) notes that IPF algorithm can be formulated as a log-linear iterative model fitting problem, noting however that if the log-linear expectation function underlying this procedure is violated then the IPF estimators may be biased (Berg and Fuller, 2009 and Griffiths, 1996). Hence, although composite estimators offer promise to enhance the performance of spatial microsimulation approaches to small area estimation such as IPF, as well as many regression-based small area estimation techniques, particular attention needs to be paid to the use of such composite estimators due to the possible bias arising from such synthetic small area estimators. This issue has however never before to our knowledge been explored empirically in the literature. Griffiths (1996) introduces the problem theoretically, drawing attention to and offering an importance discussion of the general problem. Whilst this therefore provides an important touch point to the issue that work leaves many open questions and does not provide any empirical insights or conclusions around the viability, potential and specification of composite estimators in IPF frameworks in response to its important identification of the problem.

This long overdue empirical progress is the focus of the discussion presented below. In Section 2 the general SAE problem of the population mean as a target parameter via direct estimation and IPF is introduced, as is the general framework for the derivation of sample size dependent composite estimators. In Section 3 the results of a simulation study are presented and discussed, comparing the performance of the direct estimate and IPF with two composite estimators of those direct and indirect (i.e. IPF) estimators combined. We conclude our work with a general discussion in Section 4.

2. Small Area Estimation Problem of the Population Mean

This section describes the general small area estimation problem of the population mean, introducing the direct Horvitz-Thompson estimator (Horvitz and Thompson, 1952) and the synthetic IPF reweighting algorithm before moving on to outline the general framework to the composite estimation conducted in the simulations.

2.1. Notation

Let $d = 1, \dots, D$ denote the small areas for which we want to compute the small area estimates. A sample $s \subset \Omega$ of size n is drawn from the target finite population Ω of size N. $N - n$ are

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 $s = \cup_d s_d$. $\mathrm{r_d}$ denotes the non-sampled units for small area d of $N_d - n_d$ dimension.

Here we are interested in estimating the population mean $\bar{Y}_d$ of a variable *Y* for area *d* given by:

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

Where Y_{di} denotes the value of variable *Y* for unit *i* in area *d.*

2.2. Direct Estimation

A direct estimator such as the Horvitz-Thompson estimator (HT), also known as the expansion estimator, offers a well known method to estimate population target parameters. This estimator is given by (Horvitz and Thompson, 1952):

$$\hat{\bar{Y}}_d^{Direct} = N_d^{-1} \sum_{i \in s_d} w_{di} y_{di} \tag{2}$$

where $w_{di} = \pi_{di}^{-1}$ denotes the sampling weight, and π_{di} is the first-order inclusion probability of i^{th} unit from d^{th} area in s_d.

2.3. Synthetic estimator: calibrated weights under IPF algorithm

Unfortunately, when the SAE problem arises, estimator (2) may return large variability in the estimates due to the small n_d. Furthermore, it is not possible to obtain small area estimates to those small area with zero survey sample size. IPF offers a way forwards in to the small area estimation in this context. As noted above, the goal of this method is to calculate new (calibrated) weights satisfying the following calibration equation (Deville and Särndal, 1992):

$$\sum_{i \in s} w_i \boldsymbol{x}_i = \sum_{i \in \Omega_d} \boldsymbol{x}_i = T(\boldsymbol{X}_d), \tag{3}$$

where $\boldsymbol{x}_i$ denotes a vector of auxiliary information for unit *i*. This can be viewed as an optimisation problem where the calibration equation (3) is the constraint (Deville and Särndal, 1992). However, the calibration problem is here framed in the IPF context. From a computational point of view, as highlighted in Kolenikov (2014), the algorithm to perform survey weights calibration under consists of an *outer cycle* and an *inner cycle*. The first cycle checks the convergence, meaning whether the calibration equation given by (3) is satisfied, and the second one iterates over the variables used for calibration (constraint variables), reweighting the survey cases in order to fit the aggregated small area profile on those constraints. The steps can be understood as follows (Kolenikov, 2014):

1. Initialize the iteration counter $t \leftarrow 0$ and the weights as $w_i^{0,p} \leftarrow w_i$.

2. Increment the iteration counter $t \leftarrow t+1$, thus updating the weights as $w_i^{t,0} \leftarrow w_i^{t-1,p}$.
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 applications, therefore,

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

where l 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 $R = 20$ leads to satisfying indicator values. This algorithm is area-specific and therefore needs to be iterated for each small area $d = 1, \dots, D$.

The IPF-based estimator can therefore be defined by the following formula:

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

where w_{di}^* denotes the calibrated survey weight for unit i^{th} from area d^{th}.

2.4. Composite Estimators

However, estimators built under IPF may be biased. This is not unique to IPF but is an inevitable issue in all small area synthetic estimators given the nature of the estimation problem. As stressed in Rao and Molina (2015), the possible bias arising from a synthetic estimator and the large variability (small survey sample size) or non-estimation (zero sample size) of a direct estimator can be balanced by composite estimators between the two, choosing a suitable weight in the interval [0,1] for their combination.

There are a vast number of estimators in SAE that have composite form (Rao and Molina, 2015). Two flexible and commonly used sample size dependent composite estimators are outlined below with each obtained as a linear combination between the direct estimator and the IPF-based estimator described in Sections 2.2 and 2.3. These are defined as follows (Griffiths, 1996 and Drew, et al, 1982):

$$\hat{\bar{Y}}_d^{C1} = \hat{\gamma}_d^{C1}\hat{\bar{Y}}_d^{Direct} + (1-\hat{\gamma}_d^{C1})\hat{\bar{Y}}_d^{IPF},\ \ with\ \hat{\gamma}_d^{C1} = \frac{n_d}{N_d}, \tag{5}$$

$$\hat{\bar{Y}}_{\delta,d}^{C2} = \hat{\gamma}_{\delta,d}^{C2}\hat{\bar{Y}}_d^{Direct} + \left(1-\hat{\gamma}_{\delta,d}^{C2}\right)\hat{\bar{Y}}_d^{IPF}, with\ \hat{\gamma}_{\delta,d}^{C2} = \begin{cases} 1\ if\ \ \frac{n_d}{n} \geq \delta\left(\frac{N_d}{N}\right) \\ \left(\frac{1}{\delta}\right)\frac{n_d/n}{N_d/N}\ if\ \ \frac{n_d}{n} < \delta\left(\frac{N_d}{N}\right) \end{cases}, \tag{6}$$

where $\delta \geq 0$. In Section 4, the efficiency of $\hat{\bar{Y}}_{\delta d}^{C2}$ for specific values of δ is explored. Of course, $0 \leq \hat{\gamma}_d^{C1} \leq 1\ and\ \ 0 \leq \hat{\gamma}_{\delta,d}^{C2} \leq 1$. Estimator $\hat{\bar{Y}}_{\delta,d}^{C2}$ with $\hat{\gamma}_d^{C2}$ is evaluated in Pratesi and Salvati (2008) using a regression-based small area estimator, but not in the context of a composite spatial microsimulation approach. It can be seen that both composite estimators borrow strength from other small areas. In particular, $\hat{\bar{Y}}_{\delta,d}^{C2}$ depends on δ: when δ increases the effect of borrowing strength from related small areas increases, therefore 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. Simulation study

In this context, the challenge and original contribution of the simulation results presented below is to explore the viability and impact of alternatively specified composite estimators on the performance of the small area estimation relative to that of either the direct or synthetic IPF estimators alone. Performance is assessed in terms both of bias and mean squared error (MSE) – taking into account both bias and variance. The following values of δ are explored, $\delta = \left\{0.2, \frac{1}{2}, \frac{2}{3}, 0.9, 1, 1.5, 2, 2.5, 10\right\}$, with larger values of δ denoting a higher weighting to the synthetic IPF estimator relative to the direct estimator within the composite given the same small area survey sample size.

3.1. Generating the population

This simulation study is quasi design-based, as defined within the classificatory work on types of simulation approaches in small area estimation (Münnich, 2014). In particular, the universe is a finite population further generated from the 2011 Census Microdata Individual Safeguarded Sample (Office for National Statistics, 2015) generated by extracting a stratified sample with simple random sample selection in each stratum (area). This population has the following dimensions: $650 \leq N_d \leq 1000$ with $N = 247807$ and $D = 300$ small areas. This is the population from which the simulation samples are drawn. This creates a more realistic unbalanced population for the simulation and also the population dimension N facilitates the computations in the simulation which can be intense with very large populations.

The target variable *Y* in the survey data is a binary variable denoting if the survey individual reports good general health or bad general health. The covariates used as constraint variables

in the small area estimation are age and the number of individuals in the household with long-standing illness/disability. The target parameter at the small area level is therefore the proportion of people in a bad health and the true value in the population is calculated as

$$p_d = N_d^{-1} \sum_{i=1}^{N_d} Y_{di}, \tag{7}$$

where:

$$Y_{di} = \begin{cases} 0 & if\ i\ has\ good\ health \\ 1 & if\ \ i\ has\ bad\ health \end{cases}.$$

3.2. Simulation steps

The simulation follows the following four steps:

1) From the simulated population select $S = 500$ simple random samples with without replacement selection in each area d. Sample sizes are drawn in each area according to a uniform distribution as follows:

$$n_d \sim Uniform(10, 80),$$

2) Estimate small area proportions for *D=300* small areas via IPF, direct and two composite estimators as above for each sample *s*. These estimates are denoted by $\hat{p}_{ds}^{Direct}$, $\hat{p}_{ds}^{IPF}$, $\hat{p}_{ds}{}^{C1}$, and $\hat{p}_{ds,\delta=\delta^*}^{C2}$;
3) As the true values are known from the generated population, estimate the following quality measures for each area *d* for the different types of estimates. For example, for $\hat{p}_d^{IPF}$ these are given by:

Mean Squared Error

$$MSE(\hat{p}_d^{IPF}) = S^{-1} \sum_{s=1}^{S} (\hat{p}_{ds}^{IPF} - p_d)^2, \tag{8}$$

Root Mean Squared Error

$$RMSE(\hat{p}_d^{IPF}) = \sqrt{S^{-1} \sum_{s=1}^{S} (\hat{p}_{ds}^{IPF} - p_d)^2}, \tag{9}$$

Relative Root Mean Squared Error

$$RRMSE(\hat{p}_d^{IPF}) = \frac{RMSE(\hat{p}_d^{IPF})}{\hat{p}_d^{IPF}}, \tag{10}$$

Bias

$$BIAS(\hat{p}_d^{IPF}) = S^{-1} \sum_{s=1}^{S} (\hat{p}_{ds}^{IPF} - p_d), \tag{11}$$

Relative Bias

$$RB(\hat{p}_d^{IPF}) = S^{-1}\sum_{s=1}^{S}\frac{\hat{p}_{ds}^{IPF}}{p_d} - 1, \tag{12}$$

Contribution of Bias to MSE

$$CB(\hat{p}_d^{IPF}) = \frac{[BIAS(\hat{p}_{ds}^{IPF})]^2}{MSE(\hat{p}_{ds}^{IPF})}. \tag{13}$$

4) These quality measure estimates can then be averaged across small areas to provide summary statistics. In tables and figures in Section 4.3 the subscript *d* is dropped as a result.

3.3. Results

The results of the simulation study are shown in this section. As outlined above, the focus is to assess the performance of the direct, synthetic IPF and variously specified composite estimators in terms of bias and MSE as a function of n_d.

For each of the 300 small areas Figure 1 shows the relative root mean squared error ($\times$ 100) (RRMSE%) of the direct (solid line) and IPF estimates (dashed line). RRMSE% offers a useful summary measure of the performance of the estimator considering bias and variance.

Figure 1 is ordered by the survey sample size of the small areas: the small area with the lowest survey sample size is shown to the far left of Figure 1 (sample size of 10) whilst the small area with the largest survey sample size is shown to the far right of Figure 1 (sample size of 80). To aid the reader, labelling along the horizontal axis shows the small area survey sample size of each small area.

[Insert Figure 1 here]

In line with SAE literature, Figure 1 highlights, as expected, that the RRMSE% of the direct estimator decreases as the sample size increases whereas the RRMSE% of the IPF estimates, in contrast, seems to be independent of the survey sample size. As a consequence, as measured by RRMSE% there is a gain in performance of the IPF estimator over the direct estimator when the sample size is relatively small, but these performance gains gradually diminish and then disappear entirely as the sample size increases and the RRMSE% of the direct estimates gradually falls accordingly. As expected, when sample sizes are sufficiently large (around $n_d = 60$ with $f_d = \frac{n_d}{N_d} = 0.07$ in this example) then the direct estimator provides small area estimates of equivalent or similar performance to the IPF estimates according to these analyses of RRMSE%.

The aim of this paper is to explore empirically for the first time the potential for composite estimators within spatial microsimulation small area estimation techniques such as IPF. The findings presented in Figure 1 indicate that there are, in a general sense, potential performance gains from composite estimators in these small area estimation contexts. A key resulting step of the paper is next to move beyond this general finding and to instead assess empirically the more precise way in which alternative specification(s) of those composite estimators exploit that *potential* performance gain.

To analyse this issue, Table 1 compares four quality measures from the direct and IPF estimates with those from the composite estimators: composite estimator 1 as shown above in (5) as well as nine alternatively weighted specifications of composite estimator 2 as shown in (6) above. As noted above, these specifications of composite estimator 2 vary according to the value given to the key δ parameter, with larger values of δ giving a higher weighting within the composite estimator to the synthetic IPF estimator relative to the direct estimator. For each estimator four quality measures are presented in Table 1: the relative mean squared error (RMSE); the relative root mean squared error as a percentage of the estimate proportion(RRMSE%); relative bias of the estimate as a percentage of the estimate proportion (RB%); and the percentage contribution of bias to the overall mean squared error (MSE) of the estimate that takes into account in a rounded fashion both the bias and variance of the estimate (CB%). For each summary measure the values presented in Table 1 are averages of the individual values calculated for each of the 300 small areas within the simulation.

It can be seen that by using IPF, $\hat{p}^{C1}$ and $\hat{p}^{C2}$ it is possible to provide higher performance in the small area estimation compared to the direct estimates that unbiased but sometimes with high variance. The extent to which is the case varies and the bias introduced needs to be evaluated carefully: IPF and $\hat{p}^{C1}$ provide slightly biased estimates for some areas, severely biased estimates for other areas and a high bias contribution to the MSE. Given that it is based on a more complex set of weighting options, composite estimator $\hat{p}^{C2}$ gives more flexibility and, often, better performance in terms of the trade-off between bias and variability. When δ increases it can be seen that RMSE and RRMSE% gradually decrease, but at the same time that the contribution of bias to the MSE becomes larger, although in many instances CB% remains relatively small. Evaluating across Table 1 suggests that we can optimise these trade-offs and the resulting performance of the small area estimation using a composite estimator and, more specifically, when δ is in the set $\{\frac{2}{3}, 0.9, 1, 1.5, 2\}$ within the specification of composite estimator 2. Within this set of δ the composite estimators produce a relatively small MSE – and certainly smaller MSE than the direct estimator – alongside a relatively small bias contribution to the MSE as well (even if naturally in excess of the unbiased direct estimator). Further details on the RMSE and bias for each area about each composite estimator are presented in Appendix A, again ordered by growing sample size.

[Insert Table 1 here]

Whilst Table 1 provides summary averages of these metrics, Figure 2 offers a visual overview of the full distribution of RMSE values across the 300 small areas within the simulation across all the estimation approaches. It is clear that composite estimators offer performance gains compared to both the direct estimator and synthetic IPF estimator when assessing the median, interquartile range of full range of the RMSE. Some composite specifications also perform better than others. Whilst to some extent the view of what constitutes the trade-off remains a subjective judgement of the researcher, values of δ in the set $\left\{\frac{2}{3}, 0.9, 1, 1.5, 2\right\}$ appear sensible.

[Insert Figure 2 here]

Figure 3 shows the behaviour of the weights from the various composite estimators ordered by growing sample size. To aid the reader, the legend in Figure 3 is ordered according to the order that the lines appear on the figure looking from left to right. Figure 1 above shows the tendency of the IPF estimator to drift towards the direct estimator as the survey sample size in the small areas increases. As expected, Figure 3 highlights that the pace of this tendency varies across the different composite estimators dependent upon the value of δ within each composite given that this differently controls the relative weighting between the direct and synthetic IPF components.

At one extreme, it can be seen that composite estimator 2 with a δ value of 0.2 always equals the direct estimator, and can be seen as a horizontal line at value 1 across the top of Figure 3; as Table 1 demonstrates this estimator has very low bias, but relatively high variance. At the other extreme four composite estimators – composite estimator 1 plus the three composite estimate returned by estimator 2 with the largest values of δ – never converge with the direct estimator across these survey sample sizes given the greater weighting attached to the synthetic IPF estimates within these composites. This lies behind the increased levels of bias seen within these estimators in Table 1. For composite estimators in between these two extreme positions the composite estimates do reach a point where they converge with the direct estimates, with the question being the pace at which this convergence occurs as the survey sample size increases.

[Insert Figure 3 here]

Table 2 shows in further detail the absolute small area survey sample size (n_d) and the sampling fraction ($f_d = n_d/N_d$) of each of the composite estimators at the point that the composite estimator becomes weighted entirely towards the direct estimator such that the composite and direct estimator become equivalent.

[Insert Table 2 here]

Table 3 below focuses squarely on the issue of bias by validating the true population values against the different types of estimates – the direct estimates (shown in the top row), the IPF estimates (row two) and the various composite estimates (remaining rows). Two alternative forms of external validation are presented in Table 3: Spearman's rank correlation (column one) as well as linear regression between the two. Whilst the correlation offers a single summary of fit around the line of best fit the regression goes further by enabling an understanding both of the extent to which the two distributions share a common intercept at the origin – such that $\hat{\beta}_0$ equals zero – and produce a slope coefficient ($\hat{\beta}_1$) of one.

These validation measures suggest firstly that there are three estimators that validate less well against the true value: the IPF synthetic estimator, composite estimator 1, and composite estimator 2 with δ values of around 2.5 and above. Composite estimators with δ values between the range $\left\{0.2, 0.5, \frac{2}{3}, 0.9, 1\right\}$ validate well on these metrics in contrast. Spearman's rank correlation estimates return very good ranking for the estimates given by composite estimator 2 with $\delta = \left\{0.2, 0.5, \frac{2}{3}, 0.9, 1, 1.5, 2, 2.5\right\}$.

[Insert Table 3 here]

Looking back across the various analyses presented above suggests that composite estimators can lead to substantial performance gains compared to either direct or synthetic IPF estimators and that the specification of the δ value and its interaction with the small area survey sample sizes available in the survey data are key. Taken together, in these simulations the value of δ that optimises the trade-off between the direct and synthetic parts of the composite estimator – and, in turn, the trade-off between variability and bias within the composite estimation – lies in the range $\left[\frac{2}{3}, 2\right]$.

4. Discussion

This paper focuses for the first time in the literature on empirical assessments of the viability and specification of composite estimators in IPF, a widely used spatial microsimulation approach to small area estimation. This need is motivated by the fact that in a small area context direct estimators, whilst unbiased, are typically hamstrung either by large variability (where small area survey sample sizes are small) or non-viability (where there is zero sample size for a small area). Composite estimators offer the potential to optimise the trade-off between bias and variance through the well designed combination of direct and synthetic components, yet this has remained a neglected fact of spatial microsimulation research despite its clear promise.

Using simulation based on 2011 Census Microdata Individual Safeguarded Sample for the UK the paper assesses empirically the performance of alternatively specified composite estimators with both the direct and synthetic IPF estimators across a range of key statistical performance measures. The original analyses presented demonstrate for the first time in the literature that the performance of IPF small area estimation can be enhanced by the incorporation of the IPF

estimator into a composite estimator in order to more effectively trade-off the balance between variance and bias that exist within any estimation process. Of course, the variable used in the simulation study needs to be seen as an example. In real data, there may be variables less or more spatially correlated than general health status. The role of the spatial correlation and intra-class correlation in the presented estimators needs to be evaluated empirically carefully. Particularly, this is a topic of ongoing research that we are pursuing in a regression-based small area estimation framework. Moreover, in real data applications, problematic distributions may arise, such as income distributions. These are known to have outliers and being skewed. The role of outliers here needs to be investigated.

Furthermore, by sensitivity testing findings across multiple specifications of the composite estimator the analyses enhance understanding of behaviour of the key δ value that is specified in the composite measures, and how this and how this δ value interacts with the small area survey sample size. This δ controls the relative weighting of those direct and synthetic competent elements, enabling more or less borrowing of strength from other small areas within the synthetic part of the composite. Of interest is the pace at which the different values of δ specified affect the pace at which the composite estimator converges with the direct estimator as the small area survey sample size increases – and hence as the direct estimator becomes increasingly reliable – and the impacts of these differing convergence rates on the performance of each composite estimator specification. The original empirical analyses presented highlight that the choice of δ is decisive is maximising performance of such composite indicators in IPF-based indicators as measured by MSE through the trade-off of bias and variance in the composite estimator, given the small area survey sample sizes in the dataset in use. Taking into account the range of analyses presented above, a good value of δ in these simulations look to fall in the range $\left[\frac{2}{3}, 2\right]$.

It is important that data users first check the reliability of small area direct estimates, which may not be reliable for many areas in case of large-scale national sample surveys. This is due to the unplanned domains phenomenon. Thus, indirect estimators should be used. In this paper, we focused on an IPF-based estimator that may return biased estimates for some small areas. There is a literature on internal validation of small area IPF-based estimators which readers may follow, e.g. Rahman and Harding (2017). Beside this validation, we suggest that users perform an initial exploratory bias diagnostic of the small area IPF-based estimates by simply plotting the IPF estimates against the direct estimates; these are known to be design-unbiased. Summary statistics, such as the Spearman's ranking correlation coefficient, may be helpful and crucial at this stage. Unfortunately, the bias of IPF estimates may depend on many factors (e.g. the very small sample sizes, availability of covariates, spatial correlation of covariates). This is a topic of our current research in the model-based small area estimation framework, where we aim to study different scenarios. Since in real data users may not have access to a large number of auxiliary information, especially for very confidential data on income and social exclusion, the small area biases may be unavoidable. Therefore, composite estimators, in particular, $\hat{\bar{Y}}_{\delta,d}^{C2}$ with $\delta \in \left[\frac{2}{3}, 2\right]$, provide good strengths to deal with this issue. Again, considering users' needs (trade-of between variance and bias) different values of δ can be used.

We suggest to start with a moderate value of δ, as in our study, and always investigate the bias diagnostic against the direct estimates (to check the bias of the composite estimates). This type of investigation is common in regression-based small area estimation; we refer to Brown et al. (2001) for details on this.
We also would like to stress that the evaluated composite estimators are sample size dependent so the weights γ_d do not depend on the models behind IPF estimator. Other composite estimators, such as based on the mean squared error of IPF, may be constructed, but this is a topic of future research.

It is noted that study out-of-sample areas are not incorporated in the simulation, thus $n_d > 0$. One can argue that in real data some small areas may show $n_d = 0$. Whilst we note this limitation of the present its empirical contributions remain, and two responses are possible. Firstly, it is the case very often that complex large-scale surveys do contain both large variability in the direct estimates and non-zero small area sample sizes, even if this in part depends on how 'small' one defined these sub-regional geographical units to be. Secondly, survey samples size is a foundational data constraint that presents a shared challenge across all small are estimation, reflecting in a more general sense the extreme case of zero small area survey sample sizes in which the composite estimator collapses into the synthetic part only.

Taken together, this first empirical assessment in the literature of the viability and specification of composite estimators to enhance spatial microsimulation approaches to small area estimation offers valuable original insights to enhancing the performance of these key and widely used estimation methodologies. It is hoped that this both alerts analysts and practitioners to the benefits of composite estimators when conducting such work and stimulates further much needed research efforts into the conditions affecting the good specification of those composite estimators. Further interesting work could usefully explore the relevance of composite estimation to other types of spatial microsimulation approaches to small area estimation, further analyses around other specifications of composite estimator than those assessed here, and the estimation of MSE for the proposed composite estimator.

Acknowledgements
This research has been funded by the UK Economic and Social Research Council (ESRC) National Centre for Research Methods (NCRM) grant number ES/N011619/1.

Appendix A

[Insert Figure A1 here]

[Insert Figure A2 here]

References

Agresti, A. (2002). *Categorical Data Analysis. London*: John Wiley & Sons.

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.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.

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

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., Särndal, C-E and Sautory, O (1993). Generalized Raking Procedures in Survey Sampling, Journal of the American Statistical Association 88(423): 1013-1020

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association 87(418): 376-382.

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.

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.

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.

Kolenikov, S. (2014). Calibrating Survey Data Using Iterative Proportional Fitting (Raking). The Stata Journal 14, (1): 22-59.

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.

Marshall, A (2010). *Small area estimation using ESDS government surveys – An introductory guide*. Economic and Social Data Service.

Münnich, R. (2014). Small area applications: some results from a design-based view.

International Small Area Estimation conference, SAE 2014 in Poznac, Poland. (http://www.sae2014.ue.poznan.pl/presentations/SAE2014_Ralf_Munnich_c330a31c0a.pdf).

Office for National Statistics (2015). 2011 Census Microdata Individual Safeguarded Sample (Local Authority): England and Wales. [data collection]. UK Data Service. SN: 7682, http://doi.org/10.5255/UKDA-SN-7682-1.

Pratesi, M., and Salvati, N. (2008). Small Area Estimation: the EBLUP estimator based on Spatially Correlated Random Area Effects. Statistical Methods and Applications 17: 113-141.

Pratesi, M., and Salvati, N. (2016). Introduction on Measuring Poverty at Local Level Using Small Area Estimation Methods in Pratesi (Ed.) (2016). *Analysis of Poverty Data by Small Area Estimation*. Wiley.

Purcell, N. J., and Kish, L. (1980). Postcensal Estimates of Local Areas (or Domains). International Statistical Review 48: 3-18.

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.Simpson, L and Tranmer, M (2005). Combining sample and census data in small area estimates: iterative proportional fitting with standard software. The Professional Geographer 57(2): 222-234.

Ruther, M., Maclaurin, G., Leyk, S., Buttenfield, B., and Nagle, N. (2013). Validation of spatially allocated small area estimates for 1880 Census demography. Demographic Research, Volume 29, Article 22, pp. 579-616. Available at http://www.demographic-research.org/Volumes/Vl29/22/.

Singh, A and Mohl, C (1996). Understanding calibration estimators in survey sampling, Survey Methodology 22: 107-115.

Voas, D, and Williamson, P. (2000). An evaluation of the Combinatorial Optimisation approach to the creation of synthetic microdata. International Journal of Population Geography 6: 349-366.

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.

Whitworth, A., Cater, E., Ballas, D., and Moon, G. (2017). Estimating Uncertainty in Spatial Microsimulation Approaches to Small Area Estimation: a New Approach to Solving an Old Problem. Computers, Environment and Urban System 63: 50-57.

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.

Zhang, L-C., and Giusti, C. (2016). Small Area Methods and Administrative Data Integration in Pratesi (Ed.) (2016). *Analysis of Poverty Data by Small Area Estimation*. Wiley.

Tables and Figures

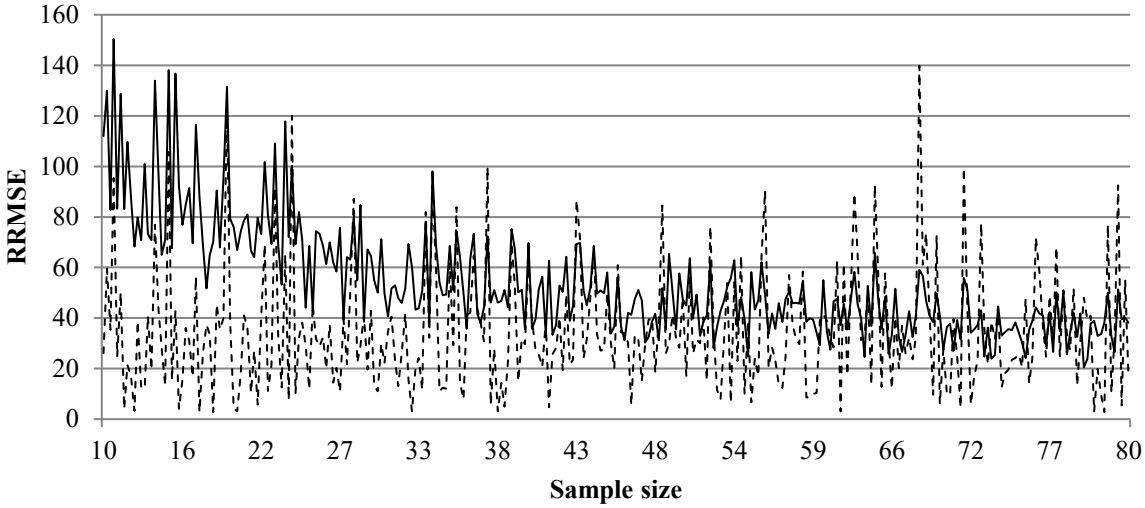

Figure 1 RRMSE% of direct estimates (__) and IPF estimates (---).

	RMSE	RRMSE%	RB%	CB%
$\widehat{\boldsymbol{p}}^{\boldsymbol{Direct}}$	0.029	54.25	-	-
$\widehat{\boldsymbol{p}}^{\boldsymbol{IPF}}$	0.018	34.45	13.18	93.22
$\widehat{\boldsymbol{p}}^{\boldsymbol{C1}}$	0.017	33.06	12.43	90.15
$\widehat{\boldsymbol{p}}^{\boldsymbol{C2}}_{\boldsymbol{\delta=0.2}}$	0.029	54.25	0.14	0.19
$\widehat{\boldsymbol{p}}^{\boldsymbol{C2}}_{\boldsymbol{\delta=0.5}}$	0.027	49.11	0.97	0.91
$\widehat{\boldsymbol{p}}^{\boldsymbol{C2}}_{\boldsymbol{\delta=2/3}}$	0.025	45.95	1.68	2.15
$\widehat{\boldsymbol{p}}^{\boldsymbol{C2}}_{\boldsymbol{\delta=0.9}}$	0.023	41.94	2.71	4.78
$\widehat{\boldsymbol{p}}^{\boldsymbol{C2}}_{\boldsymbol{\delta=1}}$	0.022	40.38	3.05	6.22
$\widehat{\boldsymbol{p}}^{\boldsymbol{C2}}_{\boldsymbol{\delta=1.5}}$	0.019	34.19	4.70	15.22
$\widehat{\boldsymbol{p}}^{\boldsymbol{C2}}_{\boldsymbol{\delta=2}}$	0.016	29.91	6.50	26.56
$\widehat{\boldsymbol{p}}^{\boldsymbol{C2}}_{\boldsymbol{\delta=2.5}}$	0.015	27.97	7.84	37.98
$\widehat{\boldsymbol{p}}^{\boldsymbol{C2}}_{\boldsymbol{\delta=10}}$	0.017	31.12	11.85	85.03

Table 1: Summary averages of RMSE, RRMSE%, RB%, and CB% from the direct, IPF and composite estimates across the 300 small areas simulated

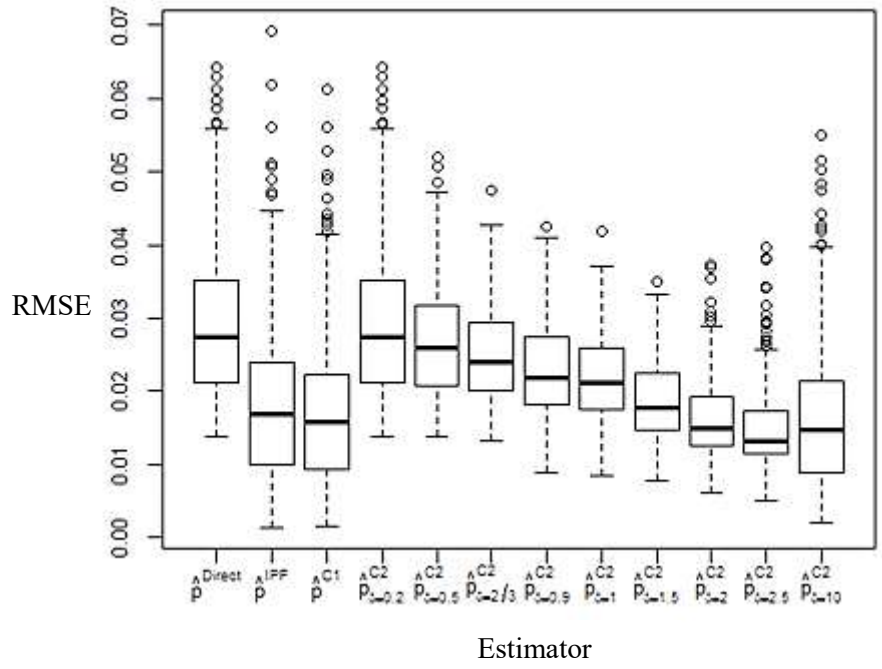

Figure 2 Box-plots of RMSE estimates.

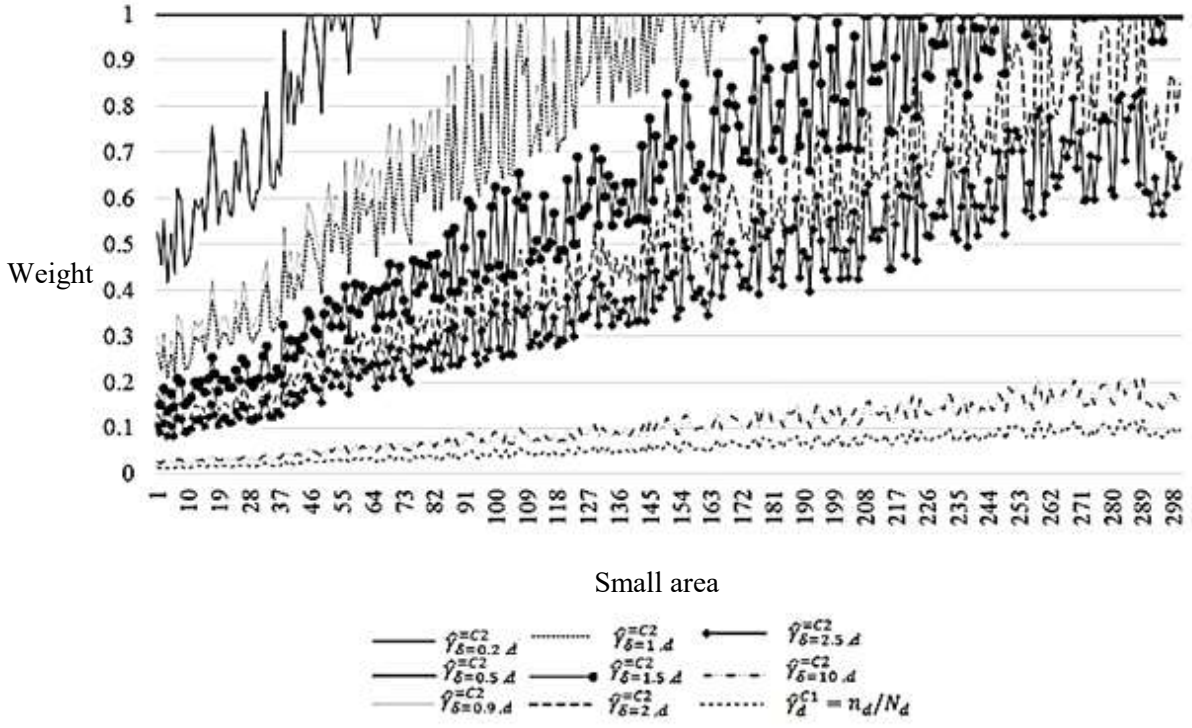

Figure 3 The behaviour of the weights within composite estimators across different small area survey sample sizes.

	$\widehat{\boldsymbol{p}}^{\boldsymbol{C1}}$		$\widehat{\boldsymbol{p}}_{\boldsymbol{\delta}}^{\boldsymbol{C2}}$							
$\boldsymbol{\delta}$	-	*0.2*	*0.5*	*2/3*	*0.9*	*1*	*1.5*	*2*	*2.5*	*10*
$\boldsymbol{n_d}$	never	always	26	36	49	49	77	never	never	never
$\boldsymbol{f_d = n_d/N_d}$			0.03	0.05	0.05	0.07	0.1			

Table 2: n_d and n_d/N_d such that the weight becomes 1 thus the composite estimator is equal to the direct estimator.

	Spearman	$\widehat{\boldsymbol{\beta}}_{\mathbf{0}}$	$\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$
$\widehat{\boldsymbol{p}}^{\boldsymbol{Direct}}$	1.00	0.00	1.00
$\widehat{\boldsymbol{p}}^{\boldsymbol{IPF}}$	0.53	0.04	0.25
$\widehat{\boldsymbol{p}}^{\boldsymbol{C1}}$	0.60	0.29	0.04
$\widehat{\boldsymbol{p}}_{\boldsymbol{\delta=0.2}}^{\boldsymbol{C2}}$	0.96	0.00	0.99
$\widehat{\boldsymbol{p}}_{\boldsymbol{\delta=0.5}}^{\boldsymbol{C2}}$	0.99	0.00	1.01
$\widehat{\boldsymbol{p}}_{\boldsymbol{\delta=2/3}}^{\boldsymbol{C2}}$	0.98	0.00	1.02
$\widehat{\boldsymbol{p}}_{\boldsymbol{\delta=0.9}}^{\boldsymbol{C2}}$	0.97	0.00	1.05
$\widehat{\boldsymbol{p}}_{\boldsymbol{\delta=1}}^{\boldsymbol{C2}}$	0.96	0.00	1.06
$\widehat{\boldsymbol{p}}_{\boldsymbol{\delta=1.5}}^{\boldsymbol{C2}}$	0.93	-0.01	1.17
$\widehat{\boldsymbol{p}}_{\boldsymbol{\delta=2}}^{\boldsymbol{C2}}$	0.90	-0.01	1.29
$\widehat{\boldsymbol{p}}_{\boldsymbol{\delta=2.5}}^{\boldsymbol{C2}}$	0.87	-0.02	1.42
$\widehat{\boldsymbol{p}}_{\boldsymbol{\delta=10}}^{\boldsymbol{C2}}$	0.65	-0.08	1.52

Table 3: External validation of direct, IPF-based and composite estimates

Figures in Appendix A

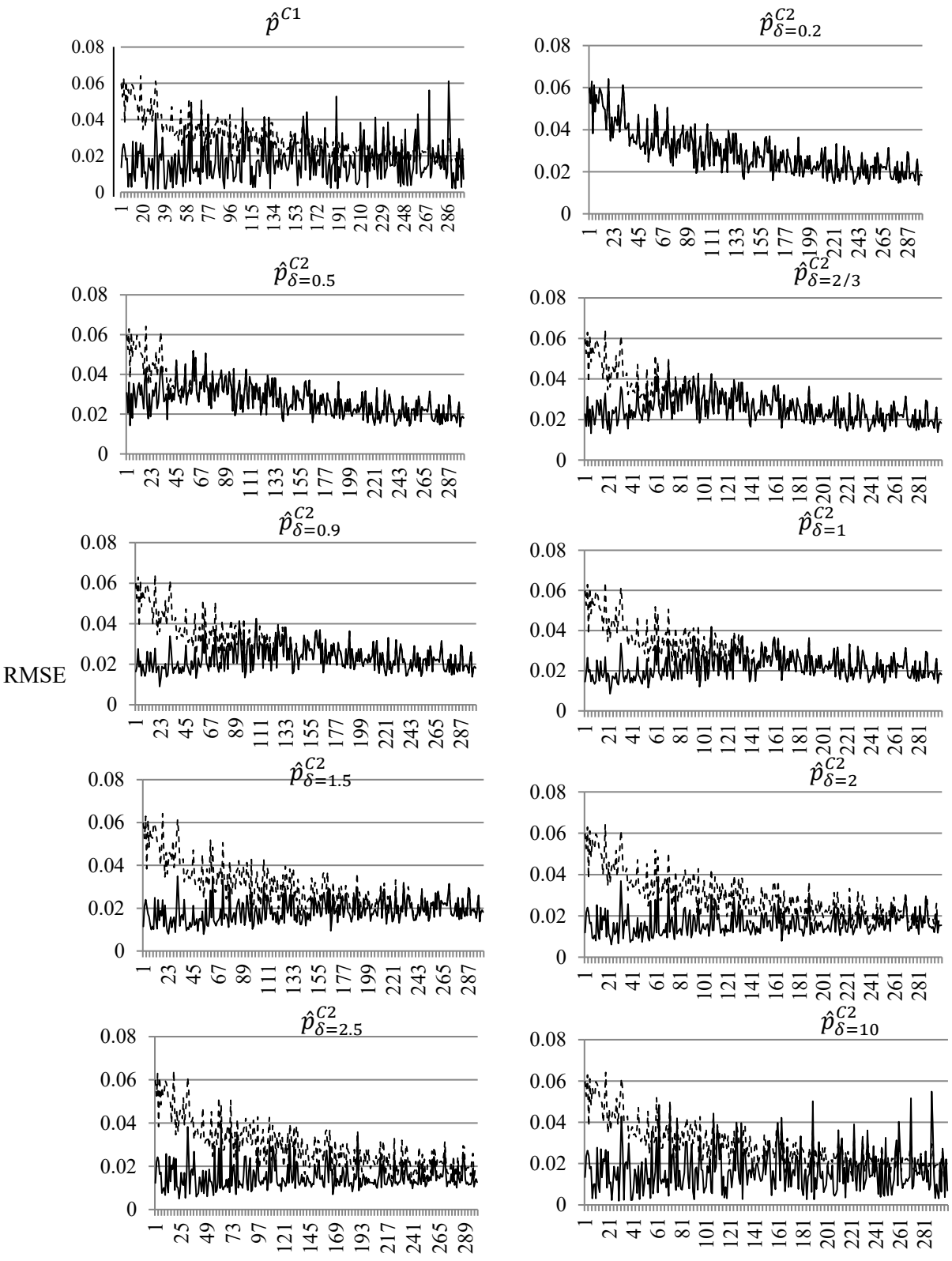

Figure A1 RMSE of composite estimates (__) and direct estimates (---).

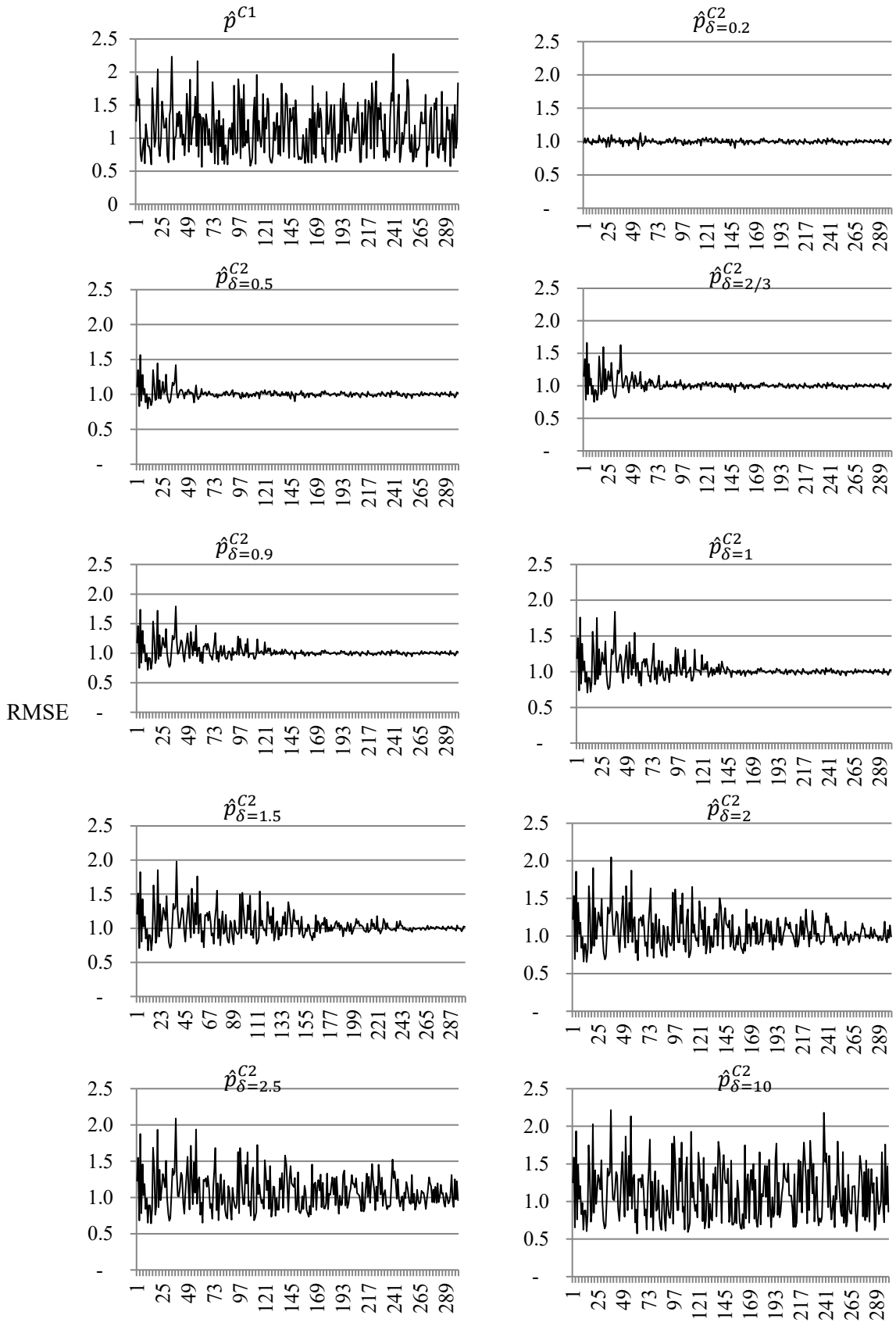

Figure A2 Ratios between composite estimates and true values.