ARTICLE IN PRESS

Computational Statistics and Data Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Small area estimation with spatial similarity

Nicholas T. Longford *

SNTL and Universitat Pompeu Fabra, R. Trias Fargas 25–27, 08005 Barcelona, Spain

ARTICLE INFO

Article history:
Received 24 July 2008
Received in revised form 8 July 2009
Accepted 7 September 2009
Available online xxxx

SUMMARY

A class of composite estimators of small area quantities that exploit spatial (distance-related) similarity is derived. It is based on a distribution-free model for the areas, but the estimators are aimed to have optimal design-based properties. Composition is applied also to estimate some of the global parameters on which the small area estimators depend. It is shown that the commonly adopted assumption of random effects is not necessary for exploiting the similarity of the districts (borrowing strength across the districts). The methods are applied in the estimation of the mean household sizes and the proportions of single-member households in the counties (comarcas) of Catalonia. The simplest version of the estimators is more efficient than the established alternatives, even though the extent of spatial similarity is quite modest.

1. Introduction

Small area estimation is concerned with inferences about quantities associated with a partition of the studied population. The population is usually a country or a region and the subpopulations its counties or districts. In most applications, the quantities of interest (targets) are means or proportions of recorded variables or of their transformations, although within-district totals, quantiles and extremes, as well as summaries of (latent) variables that are recorded subject to measurement error or another kind of distortion may also be the targets of inference. The development presented in this article is for within-district means and proportions; its extension to other quantities is outlined in Section 7. We focus first on the setting of a single variable, and in Section 5 a multivariate shrinkage adaptation is described that exploits the auxiliary information contained in other variables recorded in the same survey, similar variables recorded in other surveys or administrative registers (censuses), or in variables defined directly for the districts.

Composite estimators for small areas are defined as convex combinations of direct (unbiased) and synthetic (biased) estimators. A simple example is the composition $\tilde{\mu}_d = (1 - b_d)\hat{\mu}_d + b_d\hat{\mu}$ of the subsample mean $\hat{\mu}_d$ for the target district d and the national sample mean $\hat{\mu}$ of the target variable. The (area-specific) coefficients b_d and $1-b_d$ in this composition are set with an intent to minimise its mean squared error (MSE). The coefficients for which minimum MSE would be attained depend on some unknown parameters, which have to be estimated. As a consequence, some efficiency is lost and composition may even be counterproductive for some districts. Estimators based on empirical Bayes (EB) models have the same problem, even when a valid model is applied.

The contribution to a composite estimator for district d made by a district $d' \neq d$ depends solely on the sampling variance $\mathrm{var}(\hat{\theta}_d)$ of its direct estimator $\hat{\theta}_d$, irrespective of the distance between districts d and d'. This article develops a class of composite estimators, which address this weakness by making the contribution of each district d' to the estimator dependent on both $\mathrm{var}(\hat{\theta}_{d'})$ and its distance from the target district. The estimators combine direct estimators associated with the target

* Tel.: +34 93 542 2671; fax: +34 93 542 1746.
E-mail address: NTL@sntl.co.uk.

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.09.005

Please cite this article in press as: Longford, N.T., Small area estimation with spatial similarity. Computational Statistics and Data Analysis (2009), doi:10.1016/j.csda.2009.09.005

districts and the districts in given distances from the target. For alternative model-based solutions, see Temiyasathit et al. (2009) and Kang et al. (2009). In contrast to these and most other methods for small area estimation, we adhere to a design-based perspective, because we wish to avoid any assumptions related to underlying distributions and the functional form (linearity) of any associations. Further arguments that support this choice are presented in Section 2.1.

Section 2 introduces the setting, terminology and notation and discusses the design- and model-based perspectives. Section 3 gives details of composite estimators for the setting with no auxiliary information. Section 4 develops some refinements of the method by reusing the general idea of composition for estimating quantities that are intermediaries for small area estimation: the national mean and the between-district variance. Section 5 incorporates auxiliary information in the multivariate composite estimator, paralleling the extensions of Longford (1999, 2004). The simulation study in Section 6 compares the proposed composite estimators with their established counterparts. The estimators are for household characteristics in the counties of Catalonia. The concluding section summarises the results and discusses our experience with the proposed composite estimators.

1.1. Household size in the counties of Catalonia

Catalonia is an autonomous region of Spain (*comunitat autonoma* in Catalan), with a population of about 7 million, in about 2.5 million households, and comprises 41 counties (*comarques* in Catalan). Barcelona, which forms the county of Barcelonès, is by far the largest city in the region; it accounts for over 30% of the region's population and an even greater share of the region's economic activity. The neighbours of Barcelonès are within the city's urban sprawl and are also populous. In contrast, several counties, especially in the north and west of the region are distinctly rural and sparsely populated.

The inferential targets, the mean household size and the proportion of single-member households in each county, are of obvious interest to social scientists and the industries and services associated with residential housing. We use the results of the 2001 Spanish Census for Catalonia as the population on which we replicate the processes of sampling and estimation and empirically evaluate the MSEs of the estimators. Further background about the Census is given in Longford (2008).

In the simulations described in Section 6, we assess the gains made by assuming that counties in close proximity, and neighbouring counties in particular, have more similar summaries (profiles) of household sizes than counties located further apart and relate them to methods that disregard any spatial aspects. As a special challenge, we study the estimation for county Pla de l'Estany, which, according to the 2001 Census, has a substantially smaller average household size of 2.18 than any other county including its neighbours, even though its average in 1996 was 3.14, the highest in Catalonia. We have failed to identify any source for this discrepancy, although different administrative procedures were in place during 2001 than at earlier censuses.

The modal household size for most counties is two. The number of households with two, three and four members are similar for most counties, and the number of five-member households is several times smaller. This suggests that no familiar discrete distribution is suitable for modelling the household sizes.

We want to anticipate the precision of the small area estimators of mean household size in the counties in a future region-wide or national population survey, to inform the decision about the sampling design and to decide which estimators to apply. This we do by turning the clock back to 2001, when the last Population Census was conducted in Spain, and treating it as a sampling frame for simulated survey replicates. The distributions of household sizes in the counties are available also for 1996. We use them as auxiliary information for the estimation for 2001, mimicking the setting of a future analysis in which the direct information is from a recent population survey and the auxiliary information from a census conducted about five years earlier.

2. The setting, notation and perspectives

Suppose a population (domain or *country*) $\mathcal{P}$ is partitioned into D small areas (subdomains or *districts*) $\mathcal{P}_d$, $d = 1, \ldots, D$. We are interested in a within-district summary θ_d of a variable Y for each district d. This summary is defined by a function Θ that can be evaluated for any subpopulation of $\mathcal{P}$. Thus, $\theta_d = \Theta(\mathcal{P}_d)$. A sample $\mathcal{s}$ from $\mathcal{P}$ has a partitioning compatible with $(\mathcal{P}_1, \ldots, \mathcal{P}_D)$ into the within-district subsamples $\mathcal{s}_d = \mathcal{s} \cap \mathcal{P}_d$.

We assume that (unbiased) direct estimators $\hat\theta_d$ of θ_d, $d = 1, \ldots, D$, are defined so that they are connected by an *estimator function* $\hat\Theta$, such that $\hat\theta_d = \hat\Theta(\mathcal{s}_d)$, and that $\hat\Theta$ can be evaluated on any subsample of $\mathcal{s}$. In particular, we will evaluate $\hat\Theta$ on various unions of $\mathcal{s}_d$. Most of the results are derived for estimating the districts' population means from a survey with stratified simple random sampling, with strata coinciding with the districts. Such a design is referred to as SSRSd. The population and sample sizes of the districts are denoted by N_d and n_d, respectively. Their respective national counterparts (totals) are N and n. The within-district sampling fractions $f_d = n_d/N_d$ need not be identical. Denote $v_d = \mathrm{var}(\hat\theta_d)$ and assume that these sampling variances are known; in Section 7, we explore the impact of the uncertainty about them on the new composite estimators. For the sample means in SSRSd, $v_d = \sigma^2_{\mathrm{W},d}/n_d$, where $\sigma^2_{\mathrm{W},d}$ is the variance of Y in district d. When $\sigma^2_{\mathrm{W},d}$ coincide, we denote their common value by σ^2_{W}. For a district with $n_d = 0$ (no data), we set v_d to a very large quantity. The estimators we develop in Sections 3–5 do not depend on $\hat\theta_d$ when $n_d = 0$, so the value of $\hat\theta_d$ is immaterial in that case.

All the expectations (and variances) introduced so far relate to replications of the sampling process. We consider also expectations (averages) over the finite set of districts. For the collection $(\theta_1, \ldots, \theta_D)$, we define their (*finite-population*)

Please cite this article in press as: Longford, N.T., Small area estimation with spatial similarity. Computational Statistics and Data Analysis (2009), doi:10.1016/j.csda.2009.09.005

mean and variance as

$$\theta = \frac{1}{D}\,(\theta_1 + \cdots + \theta_D)\,, \qquad \sigma_0^2 = \frac{1}{D}\sum_{d=1}^{D}(\theta_d - \theta)^2\,,$$

respectively. In general, θ differs from the national mean $\theta^\dagger = (N_1\theta_1 + \cdots + N_D\theta_D)/N$. To avoid any confusion, we indicate the expectation with respect to *s*ampling and *d*istricts by the respective subscripts $\mathcal{S}$ and $\mathcal{D}$. Thus, $v_d = \mathrm{var}_{\mathcal{S}}(\hat\theta_d)$ and $\sigma_0^2 = \mathrm{var}_{\mathcal{D}}(\theta_d)$. For $\mathcal{S}$, the index d stands for a particular district, whereas for $\mathcal{D}$ it indicates the variable with values $d = 1, \ldots, D$. When applying $\mathcal{D}$-expectation, not only θ_d, but also n_d and v_d are treated as random variables.

2.1. Design- and model-based perspectives

We contrast two perspectives, (sampling-) design-based and model-based. In the former, there is a fixed (unchanging or *frozen*) population, with set values of all attributes, including the target variable Y and the assignment to a district, for every member of the population. At any given point of time, the populations and their divisions to districts considered in hypothetical replications of the data collection process (a survey) are identical; the sampling process is the sole source of variation.

In the model-based perspective, a (linear) model is formulated for Y as outcome in terms of some covariates X and with district-specific regressions:

$$\mathbf{y}_d = \mathbf{X}_d\boldsymbol{\beta} + \mathbf{Z}_d\boldsymbol{\delta}_d + \boldsymbol{\varepsilon}_d, \tag{1}$$

where $\mathbf{y}_d$ is the $n_d \times 1$ vector of outcomes, $\mathbf{X}_d$ the regression design matrix and $\mathbf{Z}_d$ the district-level variation design matrix for district $d = 1, \ldots, D$, $\boldsymbol{\beta}$ the vector of regression coefficients, $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_D$ a random sample from a centred multivariate normal distribution, and the $n = n_1 + \cdots + n_D$ elements of $\boldsymbol{\varepsilon}_d$ are a random sample from a centred univariate normal distribution; $\boldsymbol{\delta}_d$ and $\boldsymbol{\varepsilon}_d$ are mutually independent; see Goldstein (2002) and Longford (1993) for further background. The adaptation of (1) to generalised linear models involves a conditional distribution and a link function that connects the conditional expectation $\mathrm{E}(\mathbf{y}_d \mid \boldsymbol{\delta}_d\,; \mathbf{X}_d\,, \mathbf{Z}_d)$ to the conditional linear predictor $\mathbf{X}_d\boldsymbol{\beta} + \mathbf{Z}_d\boldsymbol{\delta}_d$; see Pinheiro and Bates (2000) and Nelder et al. (2006).

In the model-based perspective, a random (fresh) set of districts is drawn in every replication, each with a fresh set of subjects, but their values of $\mathbf{X}_d$ and the sample sizes n_d are fixed. If a subject happened to appear in two replications he or she might not be in the same district, is likely to have different values of the covariates, and the outcome will be subject to a freshly drawn deviation ε. Such a scheme is neither natural nor tenable when we seek inferences about specific (labelled) districts.

Assuming that the sampling design is well described and perfectly implemented, we regard the design-based perspective as the correct one. The model-based perspective, even with a carefully selected model, is at best an analyst's construct, because the complex processes that generate the studied population could not be credibly incorporated in a statistical model. However, in the past, design-based methods for estimating quantities associated with many subdomains (districts) have proved to be inefficient because they fail to take advantage of the similarity of the districts. This void has been filled by methods that enable *borrowing strength* across districts, as originally proposed by Efron and Morris (1972). Motivated by James and Stein (1961), Fay and Herriot (1979) obtained the same effect by applying shrinkage. The methods improve the estimation, especially for districts with very small sample sizes in the survey.

According to Longford (2007) the replication scheme, related to the status of the districts as fixed or random units, is not ignorable. The standard errors of estimators, derived from the model in (1) are, in the design-based perspective, approximately unbiased only for districts with deviations $\delta_d \doteq \pm\sigma_0$. A more subtle cause of nonignorability of the sampling design, which cannot be corrected by the traditional approaches, such as weighting, was discovered by McCullagh (2008). In the context of mixed logistic regression, he showed that the targets of inference in prediction differ when the values of a covariate are passively observed on units and when the values are assigned to them. Admittedly, this difference does not arise in the model in (1) with the normality assumptions, but it does in its adaptation to logistic regression, and also very likely in other generalised linear mixed models.

Maximum likelihood (ML) estimators, on which the model-based methods rely, are efficient only asymptotically, yet some aspects of small area estimation involve small effective samples. For example, if the value of each (univariate) deviation δ_d in (1) for a country with D districts were known, their variance $\sigma_0^2 = \mathrm{var}_{\mathcal{D}}(\delta_d)$ would have an estimator with a scaled χ^2 distribution with $D - 1$ degrees of freedom; that is, $(D-1)\hat\sigma_0^2/\sigma_0^2$ would have χ^2_{D-1} distribution. With finite subsample sizes n_d there is less information about σ_0^2, and an efficient estimator or σ_0^2 is associated with fewer degrees of freedom. The estimators of σ_0^2 applied in Section 6 to data from 41 Catalan counties, with total sample size $n \doteq 11\,600$, have approximately χ^2 distributions with only 5 – 6 degrees of freedom. For a detailed discussion and approximations, see Longford (2000) and, in a more general context, Potthoff et al. (1992).

We define the MSE of an estimator $\hat\theta_d$ for target θ_d as

$$\mathrm{MSE}\left(\hat\theta_d; \theta_d\right) = \mathrm{E}_{\mathcal{S}}\left\{\left(\hat\theta_d - \theta_d\right)^2 \,\middle|\, \theta_d\right\}; \tag{2}$$

that is, we regard each θ_d as a fixed quantity, as it is in the design-based perspective, even when it is a random quantity in the model. The conventional model-based estimators of the MSEs of small area estimators are biased. One source of bias is our

Please cite this article in press as: Longford, N.T., Small area estimation with spatial similarity. Computational Statistics and Data Analysis (2009), doi:10.1016/j.csda.2009.09.005

failure to account for uncertainty about some of the (global) parameters involved in $\hat{\theta}_d$; see Rao (2003) for addressing this problem. Another is the assumption of random effects, which from the design-based perspective is not valid. We illustrate this problem on the following example.

Suppose districts 1 and 2 have identical sample sizes in a SSRSd, the districts have a common variance σ_W^2, and the values of the target variable Y on the subjects in the sample are the sole information that is available. The model-based estimators of their population means θ_d, $d = 1, 2$, are

$$\tilde{\theta}_d = \left(1 - \hat{b}_d\right)\hat{\theta}_d + \hat{b}_d\hat{\theta}, \tag{3}$$

where $\hat{b}_d = 1/(1 + n_d\hat{\omega})$. They have identical model-related MSEs, equal to

$$\mathrm{E}\left\{\left(\tilde{\theta}_d - \theta_d\right)^2\right\} = \frac{\sigma_0^2}{1 + n_d\omega}, \tag{4}$$

ignoring the uncertainty about θ and ω and, unlike in (2), not conditioning on θ_d. This expectation is over both sampling and districts, with θ_d like a goalpost that is moved at every replication. Any (estimated) adjustment for the uncertainty about θ and ω would be the same for both districts; see Prasad and Rao (1990). Suppose the mean for district 1 differs from θ to a greater extent, say, $\theta_1 = \theta + 2\sigma_0$, and for district 2 is close to it; $\theta_2 \doteq \theta$. Then $\mathrm{var}_{\mathcal{S}}(\tilde{\theta}_1) = \mathrm{var}_{\mathcal{S}}(\tilde{\theta}_2)$; but the biases of $\tilde{\theta}_1$ and $\tilde{\theta}_2$ for their respective targets θ_1 and θ_2 differ, so $\mathrm{MSE}(\tilde{\theta}_1; \theta_1) > \mathrm{MSE}(\tilde{\theta}_2; \theta_2)$. This contradicts the equality in (4). The assumption of randomness of the districts is, therefore, not innocuous.

Most model-based estimators can be expressed as compositions of a direct and another (synthetic) estimator. This motivates the direct construction of a general composite estimator (Longford, 2004, 2005, Chapters 8 and 11). A set of alternative estimators $\hat{\theta}_d^{(h)}$, $h = 0, 1, \ldots, H$, of the same target θ_d is considered, and their convex combination

$$\tilde{\theta}_d = \sum_{h=0}^{H} b_d^{(h)}\hat{\theta}_d^{(h)} \tag{5}$$

is sought with the coefficients $b_d^{(1)}, \ldots, b_d^{(H)}$ and $b_d^{(0)} = 1 - b_d^{(1)} - \cdots - b_d^{(H)}$, for which $\mathrm{MSE}(\tilde{\theta}_d; \theta_d)$ is minimised. The estimator $\hat{\theta}_d = \hat{\theta}_d^{(0)}$ is assumed to be unbiased (or to have a known bias), to ensure that the estimation problem is well posed. The estimators $\hat{\theta}_d^{(h)}$ are called *constituent* or *basis* estimators.

The estimator in (3) is a special case of $\tilde{\theta}_d$ in (5), with $H = 1$. In (3), information from outside the target district d, mediated in $\hat{\theta}$ through $\hat{\theta}_{d'}$, $d' \neq d$, is used symmetrically. Even if the sampling variances of $\hat{\theta}_{d'}$, $d' \neq d$, were identical, we would like to give more weight to the estimators $\hat{\theta}_{d'}$ for neighbouring districts than for more distant districts, to reflect a reasonable assumption that similarity declines, or merely differs, with distance.

3. Spatial similarity

Spatial similarity is a familiar feature in a variety of contexts. Neighbouring districts tend to have similar attributes and characteristics. In environmental studies, similarity of the neighbouring geographical units adds realism to the models considered; see Elliott and Wakefield (2001), Congdon (2004), Pfeffermann and Tiller (2006). Validity of the distributional assumptions (normality) in such models is often an obstacle to their principled application. We define a class of composite estimators of θ_d, which involve no distributional assumptions and which assume a natural distance-related correlation structure of the targets θ_d, $d = 1, \ldots, D$.

We assume that a distance, $\xi(d, d')$, is defined between any two districts d and d'. This function is symmetric, nonnegative, and $\xi(d, d') = 0$ only when $d = d'$. The triangular inequality is unimportant to what follows. We assume that, in addition to zero, ξ attains only a small number H of distinct values. In our development, no generality is lost by assuming that these values are the integers $1, 2, \ldots, H$, but it is essential that for each h there be many pairs of districts (d_1, d_2) for which $\xi(d_1, d_2) = h$. Then, H will have the same role as the upper limit in (5). We associate each possible distance $h = 1, \ldots, H$ with a nonnegative (district-level) covariance $\gamma_h = \mathrm{cov}_{\mathcal{D}}\{\theta_{d_1}, \theta_{d_2} \mid \xi(d_1, d_2) = h\}$. The average squared deviation between two districts that are in distance h is

$$\left(\sigma_h^2 =\right)\ \mathrm{E}_{\mathcal{D}}\left\{\left(\theta_{d_1} - \theta_{d_2}\right)^2 \mid \xi(d_1, d_2) = h\right\} = 2\left(\sigma_0^2 - \gamma_h\right),$$

where $\sigma_0^2 = \mathrm{var}_{\mathcal{D}}(\theta_d)$ is the district-level variance. Let $\boldsymbol{\Gamma} = \mathrm{var}_{\mathcal{D}}\{(\theta_1, \ldots, \theta_D)\}$; the elements of this $D \times D$ matrix are $\Gamma_{d_1,d_2} = \gamma_{\xi(d_1,d_2)}$ when $d_1 \neq d_2$ and $\Gamma_{d,d} = \sigma_0^2$ otherwise. Note that $H = 1$ corresponds to compound symmetry (no spatial similarity). Then $\boldsymbol{\Gamma} = (\sigma_0^2 - \gamma_1)\mathbf{I} + \gamma_1\mathbf{1}\mathbf{1}^\top$, where $\mathbf{I}$ is the identity matrix and $\mathbf{1}$ the vector of unities of length implied by the context. It is easy to show that $H = 1$ implies that $\gamma_1 \doteq 0$.

For every district d, we define its *h-ring* as the subpopulation of all districts in distance h from it. In particular, $\mathcal{P}_d^{(0)} = \mathcal{P}_d$. The subpopulations $\mathcal{P}_d^{(h)}$, $h = 0, 1, \ldots, H$, define a district-specific partitioning of $\mathcal{P}$ to rings around $\mathcal{P}_d$. Denote by $\mathbf{d}_d^{(h)}$ the set of districts that form $\mathcal{P}_d^{(h)}$ and by $\mathcal{S}_d^{(h)}$ the corresponding subsample. Further, let $\theta_d^{(h)} = \Theta(\mathcal{P}_d^{(h)})$ and $\hat{\theta}_d^{(h)} = \hat{\Theta}(\mathcal{S}_d^{(h)})$,

its direct (unbiased) estimator. For example, when Θ is a (population) mean, then

$$\theta_d^{(h)} = \frac{1}{N_d^{(h)}} \sum_{d' \in \mathbf{d}_d^{(h)}} N_{d'} \theta_{d'},$$

where N_d and $N_d^{(h)}$ are the respective population sizes of district d and its h-ring; $n_d^{(h)}$ is defined as the sample counterpart of $N_d^{(h)}$. Under SSRSd, the sampling variance of $\hat{\theta}_d^{(h)}$ is

$$v_d^{(h)} = \frac{1}{N_d^{(h)\,2}} \sum_{d' \in \mathbf{d}_d^{(h)}} N_{d'}^2 \, v_{d'} \, .$$

When $v_d = \sigma_W^2 / n_d$ and the sampling fractions $f_d = n_d / N_d$ coincide, this reduces to $v_d^{(h)} = \sigma_W^2 / n_d^{(h)}$.

For each district d, we consider the compositions of the direct estimators for its rings, (5), in which $b_d^{(0)} + b_d^{(1)} + \cdots + b_d^{(H)} = 1$. We study $\mathrm{MSE}(\tilde{\theta}_d \,; \theta_d)$ as a function of the vector $\mathbf{b}_d = (b_d^{(1)}, \ldots, b_d^{(H)})^\top$, and later consider arguments of $\tilde{\theta}_d = \tilde{\theta}_d(\mathbf{b}_d)$ that are estimates of $\mathbf{b}_d$. Let $\Delta\theta_d^{(h)} = \theta_d^{(h)} - \theta_d$ (Note that $\Delta\theta_d^{(0)} = 0$.). The estimators $\hat{\theta}_d^{(h)}$ are independent and $\mathrm{E}(\hat{\theta}_d^{(h)}) - \theta = \Delta\theta_d^{(h)}$. Therefore, the MSE of $\tilde{\theta}_d$ is

$$\mathrm{MSE}\left(\tilde{\theta}_d \,; \theta_d\right) = \sum_{h=0}^{H} b_d^{(h)\,2} \left(v_d^{(h)} + \Delta^2 \theta_d^{(h)}\right). \tag{6}$$

The minimum of this function of $\mathbf{b}_d$ is found by differentiation. We obtain the conditions $b_d^{(h)} = b_d^{(0)} u_d^{(h)}$, $h = 1, \ldots, H$, where $u_d^{(h)} = v_d / (v_d^{(h)} + \Delta^2 \theta_d^{(h)})$. Of course, $u_d^{(0)} = 1$. There is a unique solution $\mathbf{b}_d^*$, and its components are

$$b_d^{(h)*} = \frac{u_d^{(h)}}{u_d^{(+)}}, \tag{7}$$

where $u_d^{(+)} = 1 + u_d^{(1)} + \cdots + u_d^{(H)}$. When $n_d = 0$ and $n_{d'} > 0$ for all $d' \neq d$, $u_d^{(h)} \gg 1$ for all $h > 0$, and so $b_d^{(h)*} \doteq 0$. Then, $\tilde{\theta}_d$ in effect does not depend on $\hat{\theta}_d$.

We refer to $\tilde{\theta}_d(\mathbf{b}_d^*)$ as the *ideal* estimators. In practice, the coefficients $b_d^{(h)}$ have to be estimated. Obviously, for any estimator $\hat{\mathbf{b}}_d^*$ of $\mathbf{b}_d^*$, $\tilde{\theta}_d(\hat{\mathbf{b}}_d^*)$ is less efficient than its ideal version $\tilde{\theta}_d(\mathbf{b}_d^*)$. If the squared deviations $\Delta^2\theta_d^{(h)}$ were known, θ_d could be estimated in the class of convex combinations (5) more efficiently than by any of the basis estimators $\hat{\theta}_d^{(h)}$, $h = 0, 1, \ldots, H$. This is so, because each $\hat{\theta}_d^{(h)}$ is itself a (trivial) convex combination with $\mathbf{b}_d$ equal to $\mathbf{0}$ (for $h = 0$) or to an indicator vector, a vector comprising $H - 1$ zeros and one unity, whilst all the components of $\mathbf{b}_d^*$ are positive.

When $H = 1$ and no distance is defined for the districts, the optimal coefficients for the composition of $\hat{\theta}_d$ with $\hat{\theta}_d^{(1)} = \hat{\theta}_{-d}$, the direct estimator for the complement of district d, are $b_d^{(0)*} = 1/[1 + v_d / \{v_d^{(1)} + (\theta_d^{(1)} - \theta_d)^2\}]$ and $b_d^{(1)*} = 1 - b_d^{(0)*}$. The resulting estimator, $(1 - b_d^{(1)*})\hat{\theta}_d + b_d^{(1)*}\hat{\theta}$, is similar, but not identical to the composition of $\hat{\theta}_d$ and $\hat{\theta}$ with $b_d = (v_d - v)/\{v_d - v + (\theta_d - \theta)^2\}$ or $b_d = v_d/(v_d + \sigma_0^2)$, see (3), because the districts are accorded weights proportional to the population size in $\hat{\theta}_{-d}$ and equal weights in $\hat{\theta}$.

As an alternative, a model, usually referring to a superpopulation of districts, is adopted and its district-level variance is estimated. Because of the uncertainty about $(\theta_d - \theta)^2$, or about σ_0^2, the estimator $\tilde{\theta}_d(\hat{\mathbf{b}}_d)$, or $\tilde{\theta}_d(\hat{b}_d)$, is not optimal in either perspective (a finite set of districts or a superpopulation) and may even be less efficient than one of the basis estimators. However, the losses in efficiency for a few districts are usually far outweighed by the gains for many, and there are various adaptations to make the estimation more conservative, to avoid substantial losses for *any* of the districts. These adaptations underestimate b_d (or every element of $\mathbf{b}_d$) and may err on the side of assigning more weight to the direct estimator.

To estimate the squared deviations $\Delta^2\theta_d^{(h)}$, we adopt a similar approach as in the setting $H = 1$. For district d and distance h, we estimate $\Delta^2\theta_d^{(h)}$ from the average of the squared distances $(\hat{\theta}_{d_1} - \hat{\theta}_{d_2})^2$ for the subset of all pairs (d_1, d_2) that are in distance h. We evaluate the district-level expectation

$$U_d^{(h)} = \mathrm{E}_{\mathcal{D}}\left(\Delta^2\theta_d^{(h)}\right) = \frac{1}{N_d^{(h)\,2}} \mathrm{E}_{\mathcal{D}}\left[\left\{\sum_{d' \in \mathbf{d}_d^{(h)}} N_{d'}\,(\theta_{d'} - \theta_d)\right\}^2\right]$$

in terms of the variance σ_0^2 and the covariances $\gamma_{h'}$, $h' = 1, \ldots, H$. In Appendix A, we derive the expression

$$U_d^{(h)} = \mathbf{r}_d^{(h)\top} \boldsymbol{\Gamma}_d^{(h)} \mathbf{r}_d^{(h)} + \sigma_0^2 - 2\gamma_h, \tag{8}$$

where $\mathbf{r}_d^{(h)}$ is the vector of all the fractions $N_{d'}/N_d^{(h)}$ for districts $d' \in \mathbf{d}_d^{(h)}$, and $\boldsymbol{\Gamma}_d^{(h)}$ is the corresponding submatrix of $\boldsymbol{\Gamma}$.

Please cite this article in press as: Longford, N.T., Small area estimation with spatial similarity. Computational Statistics and Data Analysis (2009), doi:10.1016/j.csda.2009.09.005

The district-level variance σ_0^2 can be estimated from the statistic $S_0 = \sum_{d=1}^{D} \left(\hat{\theta}_d - \hat{\theta}\right)^2$ by moment matching. We have the identity

$$\mathrm{E}_{\mathcal{S}}(S_0) = \sum_{d=1}^{D} \left\{ \left(1 - 2\,\frac{q_d}{q_+}\right) v_d + v \right\} + \sum_{d=1}^{D} (\theta_d - \theta)^2\,,$$

where q_d are the coefficients in the estimator $\hat{\theta} = (q_1\hat{\theta}_1 + \cdots + q_D\theta_D)/q_+$, and q_+ is their total. Hence, the moment-matching estimator

$$\hat{\sigma}_0^2 = \frac{S_0}{D} - \frac{1}{D}\sum_{d=1}^{D} \left(1 - 2\,\frac{q_d}{q_+}\right) v_d - v\,. \tag{9}$$

Although unbiased when all the variances v_d are exact or are estimated without bias, $\hat{\sigma}_0^2$ is inefficient when v_d are in a wide range and is especially problematic when some districts are not represented in the sample. An improvement is derived in the next section.

The covariance γ_h is estimated from the squared differences between pairs of districts in distance h. Let $m_d^{(h)}$ be the number of districts in the h-ring of district d and $m_+^{(h)} = m_1^{(h)} + \cdots + m_D^{(h)}$. Then

$$\hat{\sigma}_h^2 = \frac{1}{m_+^{(h)}} \left\{ \sum_{d=1}^{D} \sum_{d' \in \mathbf{d}_d^{(h)}} \left(\hat{\theta}_d - \hat{\theta}_{d'}\right)^2 - 2\sum_{d=1}^{D} m_d^{(h)} v_d \right\} \tag{10}$$

is an unbiased estimator of the variance $\sigma_h^2 = 2(\sigma_0^2 - \gamma_h)$ of the deviations between pairs of districts that are in distance h. Hence, $\hat{\gamma}_h = \hat{\sigma}_0^2 - \frac{1}{2}\hat{\sigma}_h^2$ is an unbiased estimator of γ_h. It may attain negative values and so its version truncated at zero should be used, even if the resulting estimator is biased. When $\gamma_h > 0$ and the 'sample' size $m_+^{(h)}$ is large, the probability that $\hat{\gamma}_h < 0$ is small. That is the rationale for defining the distance coarsely, with a small number of frequently occurring values.

Estimation of θ_d now proceeds by evaluating $\hat{u}_d^{(h)}$, $h = 1, \ldots, H$, with $(\theta_d - \theta)^2$ replaced by its estimated district-level expectation based on (8), scaling $\hat{u}_d^{(h)}$ according to (7), and using the resulting vector of coefficients $\hat{\mathbf{b}}_d^*$ in place of $\mathbf{b}_d$ in (5).

4. Some refinements

This section explores improvements in estimating the parameters θ and σ_0^2, which are intermediate quantities in estimating the targets θ_d. For both parameters, we reuse the general idea of composition and combine an unbiased and a small-variance estimator.

Although the uncertainty about θ is taken into account in our derivations, its more efficient estimation is bound to be useful, especially for estimating σ_0^2. There are two obvious candidates for estimating $\theta = \mathrm{E}_{\mathcal{D}}(\theta_d)$: $\hat{\theta}^{(\mathrm{A})} = (\theta_1 + \cdots + \hat{\theta}_D)/D$ and $\hat{\theta}^{(\mathrm{B})} = (w_1\hat{\theta}_1 + \cdots + w_D\hat{\theta}_D)/w_+$, where $w_d = 1/v_d$ is the precision of $\hat{\theta}_d$ and $w_+ = w_1 + \cdots + w_D$. When every district is represented in the sample and all v_d are finite, $\hat{\theta}^{(\mathrm{A})}$ is unbiased, but a district with a large sampling variance v_d makes a large contribution to the variance

$$V^{(\mathrm{A})} = \mathrm{var}\left(\hat{\theta}^{(\mathrm{A})}\right) = \frac{1}{D^2}\sum_{d=1}^{D} v_d.$$

When a district is not represented in the survey, $\hat{\theta}^{(\mathrm{A})}$ is either not defined, or its formally defined version has a very large (or infinite) variance. In contrast, the influence of districts with large v_d in $V^{(\mathrm{B})} = \mathrm{var}(\hat{\theta}^{(\mathrm{B})})$ is reduced, but $\hat{\theta}^{(\mathrm{B})}$ is biased. We have

$$\begin{aligned} \mathrm{MSE}\left(\hat{\theta}^{(\mathrm{B})}; \theta\right) &= \frac{1}{w_+^2}\sum_{d=1}^{D} w_d^2 v_d + \left\{ \frac{1}{w_+}\sum_{d=1}^{D} w_d\,(\theta_d - \theta) \right\}^2 \\ &= \frac{1}{w_+} + \left\{ \frac{D}{w_+}\mathrm{cov}_{\mathcal{D}}\,(w_d, \theta_d) \right\}^2 = V^{(\mathrm{B})} + B^2. \end{aligned}$$

To evaluate the MSEs of the compositions $\tilde{\theta} = (1-c)\hat{\theta}^{(\mathrm{A})} + c\hat{\theta}^{(\mathrm{B})}$, we require the identity

$$\mathrm{cov}\left(\hat{\theta}^{(\mathrm{A})}, \hat{\theta}^{(\mathrm{B})}\right) = \frac{1}{Dw_+}\sum_{d=1}^{D} w_d\, v_d = \frac{1}{w_+} = V^{(\mathrm{B})},$$

assuming that the estimators $\hat{\theta}_d$ are pairwise independent. The MSE of $\tilde{\theta} = \tilde{\theta}(c)$ is

$$\mathrm{MSE}\left\{\tilde{\theta}(c); \theta_d\right\} = (1-c)^2 V^{(\mathrm{A})} + 2c(1-c)V^{(\mathrm{B})} + c^2\left(V^{(\mathrm{B})} + B^2\right),$$

Please cite this article in press as: Longford, N.T., Small area estimation with spatial similarity. Computational Statistics and Data Analysis (2009), doi:10.1016/j.csda.2009.09.005

and this quadratic function of c attains its minimum for

$$c^* = \frac{V^{(A)} - V^{(B)}}{V^{(A)} - V^{(B)} + B^2}. \tag{11}$$

The numerator can be expressed as

$$V^{(A)} - V^{(B)} = \frac{1}{D}\left\{ \mathrm{E}_{\mathcal{D}}(v_d) - \frac{1}{\mathrm{E}_{\mathcal{D}}(w_d)} \right\}.$$

Its D-multiple is the difference of the arithmetic and the harmonic means of the (positive) variances v_d, so it is nonnegative, and equal to zero only when all v_d coincide. In that case, $\hat\theta^{(A)} = \hat\theta^{(B)}$ and the value of c is immaterial. Therefore $c^* \in [0, 1]$, and $c^* = 0$ only when $v_1 = \cdots = v_D$. Further, $c^* = 1$ only when the bias B vanishes, that is, when $\mathrm{cor}_{\mathcal{D}}(w_d, \theta_d) = 0$. In practice, the bias B of $\hat\theta^{(B)}$ is not known, and its estimation has to be based on the estimates $\hat\theta_d$; $\hat\theta^{(B)} - \hat\theta^{(A)}$ is an unbiased estimator of B. An alternative approach based on an upper bound for B^2 as prior information is explored in Longford (2008).

Composition can be applied also to the estimation of the district-level variance σ_0^2. Unlike for estimating θ, we require that the district-level distribution of θ_d be symmetric. We consider two candidate statistics from which we derive basis estimators of σ_0^2 by moment matching:

$$S_{\mathrm{A}} = \sum_{d=1}^{D} \left(\hat\theta_d - \tilde\theta\right)^2, \qquad S_{\mathrm{B}} = \sum_{d=1}^{D} n_d \left(\hat\theta_d - \tilde\theta\right)^2.$$

Let $C_d = \mathrm{cov}_{\mathcal{S}}\left(\hat\theta_d, \hat\theta\right)$ and $B_\theta = \mathrm{E}_{\mathcal{S}}\left(\tilde\theta\right) - \theta$. The two estimators of σ_0^2 are obtained from expressions for $\mathrm{E}_{\mathcal{S}}(S_{\mathrm{A}})$ and $\mathrm{E}_{\mathcal{S}}(S_{\mathrm{B}})$:

$$\hat\sigma_{\mathrm{A}}^2 = \frac{S_{\mathrm{A}}}{D} - \hat B_\theta^2 - \frac{1}{D}\sum_{d=1}^{D}(\hat v_d - 2\hat C_d) \doteq \frac{S_{\mathrm{A}}}{D} - \frac{1}{D}\sum_{d=1}^{D}\hat v_d,$$

$$\hat\sigma_{\mathrm{B}}^2 = \frac{S_{\mathrm{B}}}{n} - \hat B_\theta^2 + 2\hat B_\theta\left(\hat\theta^\dagger - \hat\theta\right) - \frac{1}{n}\sum_{d=1}^{D} n_d\,(\hat v_d - 2\hat C_d)$$

$$= \frac{S_{\mathrm{B}}}{n} + \hat B_\theta^2\, c(2-c) - \frac{1}{n}\sum_{d=1}^{D} n_d\,(\hat v_d - 2\hat C_d),$$

after substituting $\hat B_\theta = c(\hat\theta^\dagger - \hat\theta)$ and, in the first line, omitting $B_\theta^2, C_1 + \cdots + C_D$ and v, which are of lower order of magnitude than S_{A}/D and $v_1 + \cdots + v_D$. When $v_d = \sigma_{\mathrm{W}}^2/n_d$ and $\hat\theta = \sum_d n_d\hat\theta_d$, some simplification takes place, as $C_d = \sigma_{\mathrm{W}}^2/n$, and so the last summation reduces to $(D-2)\sigma_{\mathrm{W}}^2$. The bias of $\hat\sigma_{\mathrm{B}}^2$ is

$$B_{\sigma_{\mathrm{B}}^2} = \frac{1}{n}\sum_{d=1}^{D} n_d\,\Delta^2\theta_d - \frac{1}{D}\sum_{d=1}^{D}\Delta^2\theta_d = \mathrm{cov}_{\mathcal{D}}\left(\frac{n_d}{n}, \Delta^2\theta_d\right).$$

Instead of selecting $\hat\sigma_{\mathrm{A}}^2$ or $\hat\sigma_{\mathrm{B}}^2$, we combine these two estimators. To avoid some complexity, we ignore the uncertainty about B^2, v, v_d and C_d, which is insubstantial in relation to the uncertainty about $\Delta^2\theta_d$. Both estimators have the form $\hat\sigma_0^2 = \hat{\boldsymbol\psi}^\top\mathbf{G}\hat{\boldsymbol\psi} - e$ for $\hat{\boldsymbol\psi} = (\hat\theta_1, \ldots, \hat\theta_D)^\top$, a symmetric matrix $\mathbf{G}$ and a constant e. For $\hat\sigma_{\mathrm{A}}^2$, $\mathbf{G}_{\mathrm{A}} = D^{-1}(\mathbf{I} - \mathbf{1}\mathbf{q}^\top)(\mathbf{I} - \mathbf{q}\mathbf{1}^\top)$ and for $\hat\sigma_{\mathrm{B}}^2$, $\mathbf{G}_{\mathrm{B}} = (\mathbf{I} - \mathbf{1}\mathbf{q}^\top)\mathbf{R}(\mathbf{I} - \mathbf{q}\mathbf{1}^\top)$, where $\mathbf{R}$ is the diagonal matrix with the vector $\mathbf{r} = (n_1/N_1, \ldots, n_D/N_D)^\top$ on its diagonal; $\mathbf{R} = \mathrm{diag}(\mathbf{r})$.

In Appendix B, assuming symmetry of the sampling distribution of each $\hat\theta_d$, the following expression is derived:

$$\mathrm{var}\left(\hat{\boldsymbol\psi}^\top\mathbf{G}\hat{\boldsymbol\psi}\right) = \sum_{d=1}^{D} G_{dd}^2(\kappa_d - 3)v_d^2 + 2\mathbf{v}^\top\mathbf{G}^2\mathbf{v} + 4\boldsymbol\psi^\top\mathbf{G}\mathbf{V}\mathbf{G}\boldsymbol\psi,$$

where $\mathbf{v}$ is the vector of the variances v_d, $\mathbf{V} = \mathrm{diag}(\mathbf{v})$, $\boldsymbol\psi = (\theta_1, \ldots \theta_D)^\top$ and κ_d the kurtosis of the sampling distribution of $\hat\theta_d$. A similar expression is derived for $\mathrm{cov}(\hat{\boldsymbol\psi}^\top\mathbf{G}_{\mathrm{A}}\hat{\boldsymbol\psi}, \hat{\boldsymbol\psi}^\top\mathbf{G}_{\mathrm{B}}\hat{\boldsymbol\psi})$. They yield the following ideal coefficient of $\hat\sigma_{\mathrm{B}}^2$ in the composition of $\tilde\sigma^2 = (1-c^*)\hat\sigma_{\mathrm{A}}^2 + c^*\hat\sigma_{\mathrm{B}}^2$:

$$c^* = \frac{\mathrm{var}\left(\hat{\boldsymbol\psi}^\top\mathbf{G}_{\mathrm{A}}\hat{\boldsymbol\psi}\right) - \mathrm{cov}\left(\hat{\boldsymbol\psi}^\top\mathbf{G}_{\mathrm{A}}\hat{\boldsymbol\psi}, \hat{\boldsymbol\psi}^\top\mathbf{G}_{\mathrm{B}}\hat{\boldsymbol\psi}\right)}{\mathrm{var}\left\{\hat{\boldsymbol\psi}^\top(\mathbf{G}_{\mathrm{A}} - \mathbf{G}_{\mathrm{B}})\hat{\boldsymbol\psi}\right\} + B^2_{\sigma_{\mathrm{B}}^2}} = \frac{c_{\mathrm{nu}}}{c_{\mathrm{de}}},$$

where

$$c_{\mathrm{nu}} = \sum_{d=1}^{D}(\kappa_d - 3)G_{\mathrm{A},dd}\left(G_{\mathrm{A},dd} - G_{\mathrm{B},dd}\right)v_d^2 + 2\mathbf{v}^\top\mathbf{G}_{\mathrm{A}}(\mathbf{G}_{\mathrm{A}} - \mathbf{G}_{\mathrm{B}})\mathbf{v} + 4\boldsymbol\psi^\top\mathbf{G}_{\mathrm{A}}\mathbf{V}(\mathbf{G}_{\mathrm{A}} - \mathbf{G}_{\mathrm{B}})\boldsymbol\psi$$

$$c_{\mathrm{de}} = \sum_{d=1}^{D}(\kappa_d - 3)\left(G_{\mathrm{A},dd} - G_{\mathrm{B},dd}\right)^2 v_d^2 + 2\mathbf{v}^\top(\mathbf{G}_{\mathrm{A}} - \mathbf{G}_{\mathrm{B}})^2\mathbf{v} + 4\boldsymbol\psi^\top(\mathbf{G}_{\mathrm{A}} - \mathbf{G}_{\mathrm{B}})\mathbf{V}(\mathbf{G}_{\mathrm{A}} - \mathbf{G}_{\mathrm{B}})\boldsymbol\psi + B^2_{\sigma_{\mathrm{B}}^2}. \tag{12}$$

Please cite this article in press as: Longford, N.T., Small area estimation with spatial similarity. Computational Statistics and Data Analysis (2009), doi:10.1016/j.csda.2009.09.005

For normally distributed $\hat{\theta}_d$, $\kappa_d = 3$, so the summations in c_{nu} and c_{de} both vanish. By underestimating the ratio c^*, we tend to err on the side of the unbiased estimator of $\hat{\sigma}_{\text{A}}^2$ of σ_0^2. This is preferable to overestimating c^*, which exposes us, in principle, to the risk of unlimited squared bias $B^2_{\sigma_{\text{B}}^2}$. The terms involving $\boldsymbol{\psi}$ in (12) are estimated elementwise naively, using the composite estimator of $\boldsymbol{\psi}$. An alternative is to replace $\boldsymbol{\psi}\boldsymbol{\psi}^\top$ by its $\mathcal{D}$-expectation matrix implied by the parameter estimates $\hat{\sigma}_0^2, \hat{\gamma}_1, \ldots, \hat{\gamma}_D$. Composition can also be applied to estimate v_d and $\Delta^2\theta_d^{(h)}$; see Longford (2008) for details.

4.1. MSE estimation

Estimation of $\text{MSE}(\tilde{\theta}_d^{(h)}; \theta_d)$ is complicated because of its dependence on the squared deviations $\Delta^2\theta_d^{(h)}$, which are poorly estimated when the variances v_d and $v_d^{(h)}$ are large. The MSE of the ideal composition can be estimated by substituting the estimates of the coefficients (7) in (6). This yields an (approximately) unbiased estimator of the ideal MSE, and therefore an underestimate of the MSE of $\tilde{\theta}_d(\hat{\mathbf{b}}_d)$ for a district with $|\delta_d| = \sigma_0$. The estimator is always positive.

In the design-based perspective, this is the same approach as for estimating the MSE of an EB estimator, and the two types of estimators have the same deficiency of substituting $\hat{\sigma}_0^2$ for $\hat{\Delta}^2\theta_d^{(h)}$. In the simulations in Section 6, the extent of underestimation is much smaller than the bias due to $|\delta_d| \neq \sigma_0$. Instead of the estimate $\hat{\Delta}^2\theta_d^{(h)}$, we can substitute some plausible values of $\Delta^2\theta_d^{(h)}$ to obtain a range of plausible values of the MSE.

As we estimate D quantities, one per district, comparing two sets of estimators entails summarising D pairwise comparisons of MSEs, say, $m_d^{(\text{A})} = \text{MSE}(\tilde{\theta}_d^{(\text{A})}; \theta_d)$ and $m_d^{(\text{B})} = \text{MSE}(\tilde{\theta}_d^{(\text{B})}; \theta_d)$, $d = 1, \ldots, D$. In Section 6, we evaluate three summaries for this purpose: the geometric mean of the root-MSE ratios,

$$r^{(\text{AB})} = \exp\left\{\frac{1}{2D}\sum_{d=1}^{D}\log\left(\frac{m_d^{(\text{A})}}{m_d^{(\text{B})}}\right)\right\},$$

the arithmetic mean of the root-MSE differences,

$$f^{(\text{AB})} = \frac{1}{D}\sum_{d=1}^{D}\left(\sqrt{m_d^{(\text{A})}} - \sqrt{m_d^{(\text{B})}}\right),$$

and the number of districts for which $m_d^{(\text{A})} < m_d^{(\text{B})}$, denoted by $\#^{(\text{AB})}$. The first two indices compare the average efficiency of one set of estimators (A) to another (B), and $\#^{(\text{AB})}$ indicates how uniformly superior one set is to another. The summary $f^{(\text{AB})}$ is strongly influenced by the largest MSEs (smallest districts), for which a relatively small difference (say, in percentage terms) converts to a substantial difference on the linear scale of the root-MSEs. This influence is much less pronounced in $r^{(\text{AB})}$; that is the rationale for comparing the root-MSEs on the multiplicative scale. The standard deviation $s^{(\text{AB})} = \sqrt{\text{var}_{\mathcal{D}}\left(\sqrt{m_d^{(\text{A})}} - \sqrt{m_d^{(\text{B})}}\right)}$, in conjunction with $f^{(\text{AB})}$, is an alternative to $\#^{(\text{AB})}$.

5. Multivariate composition

Suppose auxiliary information is available in the form of vectors of district-level summaries $\mathbf{x}_d$, $d = 1, \ldots, D$, and their national version $\mathbf{x}$. The components of $\mathbf{x}_d$ and $\mathbf{x}$ may be direct estimators of the means or proportions of variables other than the target variable Y, obtained from the same or one or several other surveys. The components of $\mathbf{x}_d$ and $\mathbf{x}$ may also be population quantities obtained from censuses or administrative registers, or may be defined for the districts directly. There are obvious advantages if these variables are closely related to and highly correlated with Y. We require that these quantities, regarded as functions of subsamples, be well defined for every non-empty ring $\mathcal{P}_d^{(h)}$, for which they are denoted by $\mathbf{x}_d^{(h)}$.

Let $\hat{\boldsymbol{\theta}}_d = \begin{pmatrix}\hat{\theta}_d \\ \mathbf{x}_d\end{pmatrix}$, $\hat{\boldsymbol{\theta}} = \begin{pmatrix}\hat{\theta} \\ \mathbf{x}\end{pmatrix}$, $\boldsymbol{\theta}_d = \text{E}_{\mathcal{S}}(\hat{\boldsymbol{\theta}}_d)$, $\boldsymbol{\theta} = \text{E}_{\mathcal{S}}(\hat{\boldsymbol{\theta}})$, $\mathbf{V}_d = \text{var}_{\mathcal{S}}(\hat{\boldsymbol{\theta}}_d)$ and $\mathbf{V} = \text{var}_{\mathcal{S}}(\hat{\boldsymbol{\theta}})$. We assume that the variance matrices $\mathbf{V}_d$ and $\mathbf{V}$ are finite. We are concerned with estimating $\theta_d = \boldsymbol{\theta}_d^\top\mathbf{e}$, where $\mathbf{e} = (1, 0, \ldots, 0)^\top$ is the indicator of θ_d, the first component of $\boldsymbol{\theta}_d$. The derivations that follow apply for any vector $\mathbf{e}$.

We define the *ideal* estimator of θ_d as the (multivariate) composition

$$\tilde{\theta}_d = \sum_{h=0}^{H}\mathbf{b}_d^{(h)\top}\hat{\boldsymbol{\theta}}_d^{(h)}, \tag{13}$$

with the vectors $\mathbf{b}_d^{(h)}$, $h = 0, \ldots, H$, for which $\text{MSE}(\tilde{\theta}_d; \theta_d)$ is minimised, subject to the constraint that $\mathbf{b}_d^{(0)} + \mathbf{b}_d^{(1)} + \cdots + \mathbf{b}_d^{(H)} = \mathbf{e}$. The optimal vectors of coefficients $\mathbf{b}_d^{(h)}$ are found by differentiating the MSE

$$\text{MSE}\left(\tilde{\theta}_d; \theta_d\right) = \sum_{h=0}^{H}\mathbf{b}_d^{(h)\top}\left(\mathbf{V}_d^{(h)} + \Delta\boldsymbol{\theta}_d^{(h)}\Delta\boldsymbol{\theta}_d^{(h)\top}\right)\mathbf{b}_d^{(h)},$$

Please cite this article in press as: Longford, N.T., Small area estimation with spatial similarity. Computational Statistics and Data Analysis (2009), doi:10.1016/j.csda.2009.09.005

Table 1
Comparisons of the sets of small area estimators. Quantities $r^{(AB)}$ and $[\#^{(AB)}]$ below and $f^{(AB)}$ and $(s^{(AB)})$ above the diagonal. The quantities are defined in Section 4.1.

	Direct	*U-Comp-1*	*U-Comp-2*	*U-Comp-3*	*B-Comp-1*	*B-Comp-2*
Direct		0.0804 (0.0996)	0.0904 (0.1111)	0.0881 (0.1130)	0.0965 (0.1256)	0.1028 (0.1280)
U-Comp-1	0.634 [39]		0.0101 (0.0219)	0.0077 (0.0308)	0.0162 (0.0377)	0.0224 (0.0521)
U-Comp-2	0.546 [35]	0.861 [29]		−0.0023 (0.0123)	0.0061 (0.0339)	0.0123 (0.0438)
U-Comp-3	0.539 [34]	0.851 [27]	0.988 [18]		0.0084 (0.0386)	0.0146 (0.0450)
B-Comp-1	0.533 [39]	0.841 [37]	0.977 [22]	0.989 [22]		0.0000 (0.0243)
B-Comp-2	0.466 [39]	0.735 [37]	0.853 [31]	0.864 [28]	0.874 [35]	

where $\Delta\boldsymbol{\theta}_d^{(h)} = \boldsymbol{\theta}_d^{(h)} - \boldsymbol{\theta}_d$. We obtain the conditions

$$\mathbf{b}_d^{(h)} = \left(\mathbf{V}_d^{(h)} + \Delta\boldsymbol{\theta}_d^{(h)}\Delta\boldsymbol{\theta}_d^{(h)\top}\right)^{-1}\mathbf{V}_d\mathbf{b}_d^{(0)}$$

and the unique solution

$$\mathbf{b}_d^{(h)*} = \left\{\mathbf{I} + \sum_{h'=1}^{H}\left(\mathbf{V}_d^{(h')} + \Delta\boldsymbol{\theta}_d^{(h')}\Delta\boldsymbol{\theta}_d^{(h')\top}\right)^{-1}\mathbf{V}_d\right\}^{-1}\left(\mathbf{V}_d^{(h)} + \Delta\boldsymbol{\theta}_d^{(h)}\Delta\boldsymbol{\theta}_d^{(h)\top}\right)^{-1}\mathbf{V}_d\mathbf{e},$$

if each matrix inverse exists. A sufficient condition for these inverses to exist is that each $\mathbf{V}_d^{(h)}$ is non-singular. A special case of $\mathbf{b}_d^{(h)*}$, when there is no auxiliary information, is the univariate solution given by (7).

The vectors $\mathbf{b}_d^{(h)*}$ have to be estimated. This entails estimating the matrix $\Delta\boldsymbol{\theta}_d^{(h)}\Delta\boldsymbol{\theta}_d^{(h)\top}$ by moment matching or by its estimated $\mathcal{D}$-expectation, which in turn is a function of the variance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Upsilon}^{(h)}$, $h = 1, \ldots, H$, the respective multivariate counterparts of the district-level variance σ_0^2 and covariances γ_h.

The variance matrices $\mathbf{V}_d = \mathrm{E}_{\mathcal{S}}(\hat{\boldsymbol{\theta}}_d)$ and $\mathbf{V} = \mathrm{E}_{\mathcal{S}}(\hat{\boldsymbol{\theta}})$ can be estimated without bias by the multivariate version of the Yates-Grundy estimator, see Särndal et al. (1992). The estimators can be pooled when multivariate homoscedasticity is assumed. The matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Upsilon}^{(h)}$ are estimated elementwise. Their diagonal elements are estimated by the same moment-matching method as in the univariate composition. For the off-diagonal elements, the method is adapted by matching the moments of a (weighted) sample covariance matrix.

6. Empirical evaluation

We consider a survey with SSRSd and the same sampling fraction of 1/200 of households in every county of Catalonia. The targets are the within-county average household sizes. The within-county subsample sizes have binomial distributions, so that the sample sizes for the two least populous counties are zero or one with non-trivial probabilities. The direct estimators are sufficiently precise for any conceivable purpose for a few most populous counties, but they are of next to no value for the least populous counties. The simulation of the sampling and estimation processes is conducted with 500 replications.

In bivariate composite estimation, we use the population data from 1996 as the auxiliary information. It is without any sampling variation, but we nevertheless associate each county-level mean for 1996 with a token variance of 0.0001, to represent the presumed imperfection of the data. The county-level average household sizes have dropped from 1996 to 2001 by 0.10–0.34, except for Pla de l'Estany, for which the drop was by 0.96.

The comparisons of the sets of estimators are summarised in Table 1. The rows and columns of the table correspond to the sets; the cells under the diagonal contain the geometric mean of the root-MSE ratios, $r^{(AB)}$, for row A and column B, with the number of counties for which A is more efficient than B, $\#^{(AB)}$, in brackets. The cells above the diagonal contain the arithmetic mean $f^{(AB)}$ and, in parentheses, the standard deviation $s^{(AB)}$ of the differences of root-MSEs. The notation is explained below and the definitions of the estimators are listed in Table 4 in the Appendix. Using $H = 2$ corresponds to distinguishing between neighbours (distance 1) and non-neighbours, and $H = 3$ to classifying the non-neighbours as neighbours' neighbours (distance 2) and more distant counties.

The geometric means of the ratios, $r^{(AB)}$, are compatible with the ordering of the estimators from the top (*Direct*) to the bottom row (*B-Comp-2*) of the table. Thus, the univariate composition without distance similarity (*U-Comp-1*) is on average more efficient than the direct estimation by $100 \times (1 - 0.634) = 36.6\%$. The univariate composition with the distance truncated at $H = 2$ (*U-Comp-2*) is 13.9% on average more efficient than the estimation with *U-Comp-1*, and the univariate composition with the distance truncated at $H = 3$ (*U-Comp-3*) is only slightly more efficient on average than *U-Comp-2*. The bivariate composition without distance similarity (*B-Comp-1*) is only slightly more efficient than *U-Comp-3*, but the composition with the distance truncated at $H = 2$ (*B-Comp-2*) is more efficient on average than *B-Comp-1* by 12.6%.

Please cite this article in press as: Longford, N.T., Small area estimation with spatial similarity. Computational Statistics and Data Analysis (2009), doi:10.1016/j.csda.2009.09.005

Table 2
Comparisons of the sets of small area estimators for Catalonia without county Pla de l'Estany. The same layout is used as in Table 1.

	Direct	*U-Comp-1*	*U-Comp-2*	*U-Comp-3*	*B-Comp-1*	*B-Comp-2*
Direct		0.0874 (0.0940)	0.0980 (0.1017)	0.0956 (0.1020)	0.1195 (0.1142)	0.1135 (0.1131)
U-Comp-1	0.603 [39]		0.0106 (0.0156)	0.0082 (0.0223)	0.0321 (0.0331)	0.0262 (0.0420)
U-Comp-2	0.523 [35]	0.867 [28]		−0.0024 (0.0117)	0.0214 (0.0369)	0.0155 (0.0466)
U-Comp-3	0.521 [34]	0.864 [25]	0.997 [17]		0.0239 (0.0417)	0.0180 (0.0513)
B-Comp-1	0.432 [39]	0.716 [37]	0.826 [27]	0.828 [25]		−0.0059 (0.0260)
B-Comp-2	0.451 [36]	0.747 [34]	0.862 [27]	0.864 [24]	1.043 [17]	

The ordering of average efficiency implied by the summaries $f^{(AB)}$ differs from the ordering implied by $r^{(AB)}$ only by the elementary swap of the sets *U-Comp-2* and *U-Comp-3*. The average reduction of root-MSEs for *U-Comp-2* over *Direct* (0.0904) is greater than for *U-Comp-3* over *Direct* (0.0881). The arithmetic means of the root-MSE reductions require a reference to the scale of the outcome variable, whereas the geometric means $r^{(AB)}$ are on an absolute scale; 0.466 for the comparison of *B-Comp-2* with *Direct* corresponds to 53.4% average reduction of root-MSE and $100\times(1-0.466^2) = 78.3\%$ reduction of MSE.

In univariate composition, truncating the distances at $H = 3$, using *U-Comp-3*, yields an average root-MSE reduction of only 1.2% over *U-Comp-2*, and less truncation, setting $H > 3$, is counterproductive, as assessed by both $r^{(AB)}$ and $f^{(AB)}$. An improvement from $H = 2$ to $H = 3$ is recorded for only 18 counties. In bivariate composition, truncating the distances at $H > 2$ is counterproductive; the only spatial feature worth incorporating is whether two counties are neighbours or not.

The values of the parameters σ_0^2 and γ_h, obtained with precision from the census data, provide a partial explanation for the performance of the composite estimators. We have $\sigma_0^2 = 0.0244$. When $H > 1$, $\gamma_1 = 0.0109$, when $H > 2$, $\gamma_2 = 0.004\,58$, and when $H > 3$, $\gamma_3 = 0.002\,01$, so that the covariances decline about 2.3 times per unit distance. However, γ_4 and γ_5 are negative when $H > 5$. The values of γ_h do not depend on the truncation applied, as long as $H > h$. When $H = 1$, $\gamma_1 = 0$; otherwise $\gamma_H < 0$ for $H > 1$.

6.1. Pla de l'Estany and other extreme counties

The univariate composite estimators (*U-Comp-h*, $h = 1, 2, 3$) are more efficient than the direct estimators for most counties (34 – 39), and the bivariate composite estimators (*B-Comp-1* and *B-Comp-2*) are more efficient than the corresponding univariate estimators (*U-Comp-1* and *U-Comp-2*) for 37 and 31 counties (out of 41), respectively. In all these comparisons, the estimator for Pla de l'Estany is in the minority. For example, only two counties have the root-MSEs for *U-Comp-1* greater than for *Direct*, Pla de l'Estany by 63% and Vallès Occidental by only 1%. Composite estimation for Pla de l'Estany is unsuccessful because the county is a distinct outlier, and using any auxiliary information in the estimation for it is, in effect, misleading.

Even setting aside Pla de l'Estany, the root-MSE reductions of one set of estimators over another are not closely related to the average sample size. As an extreme example, the root-MSE for *U-Comp-1* is five times smaller than for the direct estimator for Val d'Aran, the county in the northwest corner of the region. The corresponding reduction for Alta Ribagorça, the least populous county, is 'only' 3.45-fold. Composite estimation for Val d'Aran is so effective, because its deviation $\theta_d - \theta = 0.03$ is very small. Models for spatial similarity, using $H > 1$, are not useful for Val d'Aran, because it has only two neighbours, both of them sparsely populated and with very different means θ_d. Montsià, the southernmost county, has only one neighbour, Ribera d'Ebre. The two counties happen to have very similar population means of household sizes (2.80 and 2.82), so the composition for Montsià with $H \geq 2$ is useful.

The results are negatively affected by Pla de l'Estany, because the county is so exceptional. Table 2 summarises the estimators applied to the 40 counties, excluding Pla de l'Estany. It shows greater average gains by the composite estimators over the direct estimators. They are greatest for *B-Comp-1* ($r^{(AB)} = 0.533$ vs. 0.432 without Pla de l'Estany). Care has to be exercised when comparing the entries in Tables 1 and 2 because all the entries are by themselves comparisons. The impact of excluding Pla de l'Estany can be described more compactly by comparing the summaries for the other 40 counties in the two analyses, one with data from Pla de l'Estany used as auxiliary information and the other without. The root-MSEs are reduced for a majority of the counties, and so are the summaries $r^{(AB)}$ and $f^{(AB)}$, by between 4% (for *U-Comp-3*) and 25% (*B-Comp-1*). The reductions are smaller with the models for spatial similarity.

6.2. MSE estimation

The root-MSEs of the six sets of estimators are compared for the individual counties in the pairwise plot in Fig. 1. Each off-diagonal panel has the same scale, with the counties in the ascending order of their population sizes on the horizontal axis and the root-MSEs for each set of estimators on the vertical axis. Further details are given in the figure caption. The reductions of the MSE from direct to univariate composite estimators are substantial for the less populous counties and are

Please cite this article in press as: Longford, N.T., Small area estimation with spatial similarity. Computational Statistics and Data Analysis (2009), doi:10.1016/j.csda.2009.09.005

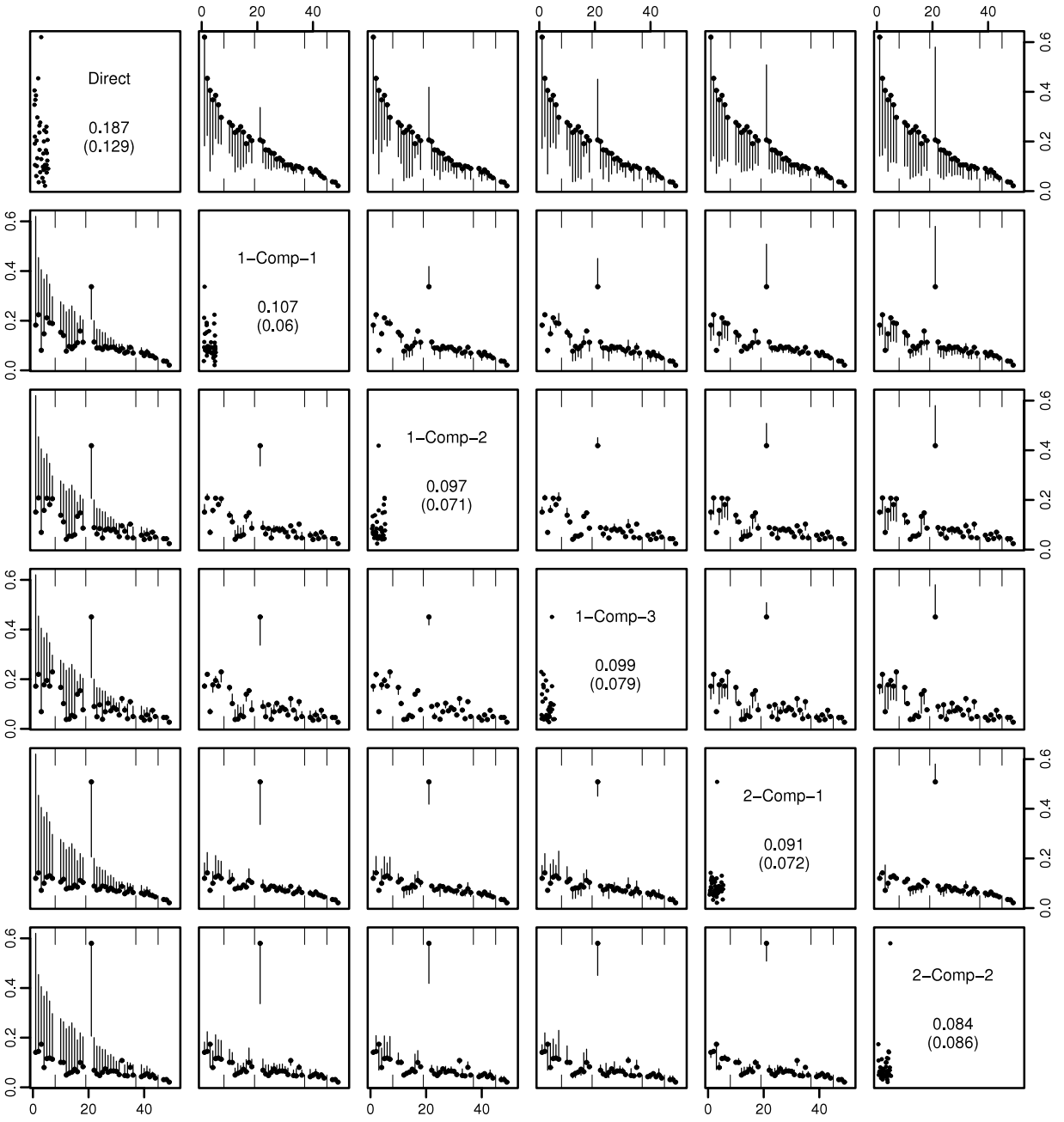

Fig. 1. The empirical root-MSEs of the small area estimators of the mean household sizes in the counties of Catalonia. Each vertical line connects the root-MSEs of estimator A (row) with the estimator B (column) for a county. The root-MSEs for A are marked by filled circles. The counties on each horizontal axis are in the ascending order of population size, with gaps and ticks placed at the top and bottom of each panel at population sizes 4000, 10 000, 44 000 and 100 000. The diagonal panels list the means and standard deviations of the root-MSEs and plot the root-MSEs at their left-hand margins, spread horizontally at random to avoid extreme overprinting.

more modest from univariate to bivariate composite estimators. There are a few reversals, but the poor performance of all the composite estimators for Pla de l'Estany stands out.

From a single sample, the MSEs of the composite estimators $\tilde{\theta}_d(\hat{\mathbf{b}}_d^*)$ can be estimated naively, as minima of the MSEs of the corresponding ideal estimators $\theta(\mathbf{b}_d^*)$, with the squared deviations $(\theta_d - \theta)^2$ or $(\theta_d - \theta_d^{(h)})^2$ replaced by $\hat{\sigma}_B^2$. Composition can be applied also to estimate these MSEs; see Longford (2007) for details. However, the application of this method is feasible only for those estimators that ignore the distance.

Fig. 2 compares the empirical (simulation-based) and analytical (data-based) estimators of the root-MSEs of the univariate composite estimators with the distance ignored and with distance used with a truncation at $H = 2$. The counties

Please cite this article in press as: Longford, N.T., Small area estimation with spatial similarity. Computational Statistics and Data Analysis (2009), doi:10.1016/j.csda.2009.09.005

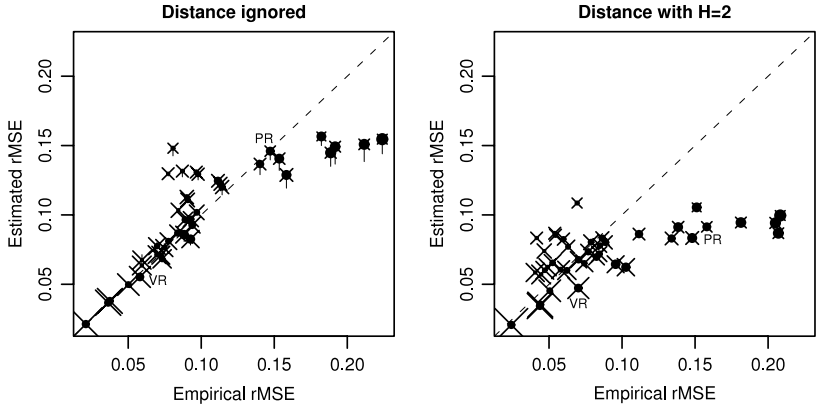

Fig. 2. The empirical and data-based estimators of the root-MSE of the univariate composite estimators. Filled circles indicate the size of the absolute deviation $|\theta_d - \theta|$, and crosses the population size of the county. The values for counties Priorat and Vallès Oriental are marked by their respective acronyms, PR and VR.

are represented in both panels by filled circles (•) with diameters linearly related to the absolute deviations $|\theta_d - \theta|$ and crosses (×) of size linearly related to the fourth root of the population sizes. In the left-hand panel, we have two sets of data-based estimators; one estimates each $\Delta^2\theta_d$ by $\hat{\sigma}_0^2$ (averaging), and the other estimates each $\Delta^2\theta_d$ by composition. They are connected by vertical segments and the filled circles are placed at the latter values. County Pla de l'Estany is omitted from both panels; its root-MSEs are grossly underestimated.

The diagram shows that the estimators of the root-MSEs tend to be positively biased for counties with small absolute deviations and negatively biased for counties with large absolute deviations. Estimation of the root-MSE is more precise for more populous counties. Underestimation is substantial for several counties, which have small population sizes and large absolute deviations.

Estimation of the root-MSEs entails much less bias when distance is ignored than when it is taken into account. When the distance is ignored, replacing $(\theta_d - \theta)^2$ with σ_0^2 entails no 'error' only when $|\theta_d - \theta| \doteq \sigma_0$. Two such counties, Priorat (PR) and Vallès Oriental (VR), are marked in the diagram. When the distance is ignored, the root-MSEs are estimated with only slight bias, whereas with $H = 2$ they are underestimated substantially. In bivariate composition and with $H > 2$, the uncertainty is even more pronounced. When the distance is ignored, the root-MSEs are estimated with much greater precision even in bivariate shrinkage. Without Pla de l'Estany, the estimators of the root-MSEs have much smaller biases. Details are omitted.

6.3. Empirical Bayes estimation

Every set of composite estimators has its counterpart set of estimators based on the EB models in which the counties are associated with random effects. Univariate composition corresponds to models with no covariates. The random effects are independent when $H = 1$ (no spatial similarity), and otherwise have the covariance matrix $\mathbf{\Gamma}$ defined in Section 3; they correspond to spatial EBLUP. With the normality assumptions, which admittedly are grossly violated, such models can be fitted by an iterative (Newton–Raphson) algorithm that maximises the corresponding log-likelihood. Conditionally on the regression and the within-county variance, they use the same sufficient statistics as their composition counterparts. We fitted these models for the counterparts of the estimators *V-Comp-h*, V = U or B, and $h = 1, 2, 3$. The sets of these EB estimators are for a majority of districts, as well as on average, less efficient than their composition counterparts. The average root-MSEs are greater by between 8% and 24%. For the most populous counties, the EB estimators are almost uniformly, although only slightly, less efficient than the composite estimators. In contrast, EB estimators are more efficient for a few sparsely populated counties, but these sets of counties differ from one model to the other.

The MSEs of the estimators are estimated from the conditional variances evaluated in the concluding iteration. They display similar features as the MSEs estimated in the composite estimation: they are approximately unbiased for 'typical' counties, overestimate the MSEs for counties with means close to the national mean and underestimate them for the counties with large deviations from the national mean.

6.4. The proportions of single households

The county-level percentages of single-member households are estimated by the same methods as the mean household sizes. The dichotomous nature of the outcome variable entails no additional complexity to the analysis of continuous variables. We found that the root-MSE reductions are in general smaller than for estimating the mean household sizes, but are nevertheless substantial. For bivariate composition, taking into account the distance truncated at $H = 2$ yields substantial gains. With a truncation at $H = 3$, the root-MSE reductions largely cancel out, although the differences between the root-MSEs for $H = 2$ and $H = 3$ are substantial for a few counties. The comparisons of the MSEs are summarised in Table 3.

Please cite this article in press as: Longford, N.T., Small area estimation with spatial similarity. Computational Statistics and Data Analysis (2009), doi:10.1016/j.csda.2009.09.005

Table 3
Comparisons of the sets of small area estimators of the proportion of single-member households in the counties of Catalonia. The same layout is used as in Table 1.

	Direct	*U-Comp-1*	*U-Comp-2*	*U-Comp-3*	*B-Comp-1*	*B-Comp-2*
Direct		0.0200 (0.0306)	0.0228 (0.0377)	0.0217 (0.0400)	0.0243 (0.0399)	0.0263 (0.0441)
U-Comp-1	0.684 [36]		0.0028 (0.0119)	0.0017 (0.0151)	0.0043 (0.0118)	0.0063 (0.0179)
U-Comp-2	0.560 [35]	0.819 [29]		−0.0011 (0.0044)	0.0015 (0.0118)	0.0035 (0.0116)
U-Comp-3	0.560 [35]	0.819 [29]	1.001 [19]		0.0026 (0.0138)	0.0046 (0.0122)
B-Comp-1	0.590 [39]	0.863 [39]	1.054 [15]	1.053 [20]		0.0002 (0.0096)
B-Comp-2	0.493 [37]	0.722 [35]	0.882 [31]	0.881 [27]	0.837 [31]	

7. Discussion

Composition in small area estimation can be broadly interpreted as a way of exploiting the similarity of the districts. When similarity is related to the distances among the districts the composition can be based on the direct estimators for the rings of the target district. Efficient inference about the extent and pattern of similarity is a key to its successful application. In distinctly non-asymptotic settings, this calls for a parsimonious model for similarity, in which uncertainty about the estimated parameters is more than offset by the improved description of similarity. This balancing act is as important as in the EB estimation. In composite estimation we do not have to associate districts with random effects, nor these effects with a distribution, which are essential elements of the EB analysis.

When distance is ignored, the composite estimators attain greater stability because direct estimators are combined only with the estimator of the overall mean. When distances are used, the basis estimators for some rings have large variances (as do many direct estimators $\hat{\theta}_d$), so composite estimators are effective only when this drawback is compensated by advantages flowing from a well-specified distance function ξ for the districts.

The magnitudes of the MSEs can be anticipated with neither EB nor composite estimators, because they depend on the targets θ_d. However, we can identify likely problems with the composition solely from the counties' neighbours and their population sizes (and other auxiliary information, such as a past census, when we intend to use it). The problems may be addressed by altering the definition of the distance.

The replacement of the various squared deviations, such as $(\theta_d - \theta_{d'})^2$ and $(\theta_d - \theta_d^{(h)})^2$, by their district-level expectations is a source of imprecision (uncertainty) in the design-based perspective. Validity of the model helps us only to attain an approximate balance of the errors due to such substitution. This issue is moot only when the similarity is perfect (e.g., when $\sigma_0^2 = 0$ or $\gamma_1 = \sigma_0^2$). The estimation of model parameters introduces another layer of uncertainty, which affects this balance. The composite estimators are nonlinear functions of these deviations, and so unbiased estimation of their district-level expectations (averages) is of mainly illusory value.

The estimation of district-level counts and totals has to be based on the estimation of the respective proportions and means, because the former are much less likely to be similar than the latter. The estimation of nonlinear summaries of the target variable Y, such as percentiles and extremes, presents considerable challenges. These can be traced through the steps in the estimation of the means and proportions: finding (approximately unbiased) direct estimators of the targets and of their sampling variances; extending them to all the rings $\mathcal{P}_d^{(h)}$; evaluating the district-level expectations (averaging); and estimating the coefficients in the composition for estimators $\hat{\theta}$ that are linear functions of $\hat{\theta}_d$. Approximations (e.g., by linearisation) may be necessary, weighing the choice of the model (the distance structure) toward parsimony. In designs other than SSRSd, the direct estimators $\hat{\theta}_d$ are correlated. Estimating means and proportions directly remains tractable, but all the expressions involving disjoint sets of districts, are more complex. In practice, these correlations are presumed to be small and are often ignored.

Uncertainty about the variances of the direct estimators, v_d, can be addressed together with the uncertainty about the district-level variance σ_0^2, because the MSEs of the composite estimators depend on the sums of their reciprocals, $1/v_d + 1/\sigma_0^2$, or their matrix versions $\mathbf{V}_d^{-1} + \mathbf{\Sigma}_0^{-1}$. Inflation or overestimation of these scalar sums (or diagonals of the matrices) injects stability. Erring on the side of positive bias in estimating v_d or ω is, therefore, less harmful than erring by the same amount in converse. Estimators of σ_0^2 are negatively associated with v_d, so overestimation of v_d is reflected by underestimation of σ_0^2. The impact of the uncertainty about v_d (or ω) can be studied empirically by replacing its estimation in the algorithm used in the replications with the (fixed) population quantity v_d, and comparing the results of the two sets of replications, before and after the replacement. We found that the uncertainty about v_d has a much weaker impact than the uncertainty about σ_0^2. Various implementations of the bootstrap, Efron and Tibshirani (1993), are effective for data-based estimation of the sampling variance of small area estimators, but they are even more difficult to adapt for the estimation of the (design-based) bias.

The EB methods are the obvious alternative to the method presented in this article. For small area estimation, the main source of bias in the MSE estimation is the model assumption of randomness of the districts, which is in conflict with the

Please cite this article in press as: Longford, N.T., Small area estimation with spatial similarity. Computational Statistics and Data Analysis (2009), doi:10.1016/j.csda.2009.09.005

design-based perspective. The uncertainty about the variance and covariance parameters contributes to the bias much less, although this contribution increases with model complexity. The distributional assumptions and the functional form of the regression in ML are an unnecessary burden for the analysis, more so that the target variables rarely have an easy-to-identify distribution, except for the binary.

Setting the details of a composite estimator presents a problem analogous to the model selection in the EB and spatial methods. Models can be compared by various information criteria, but no such framework is available for composite estimators. However, the correspondence of sets of the EB and composite estimators can be exploited by selecting an EB model, and using the corresponding composite estimators. In our simulations, model selection prefers the spatial structure with $H = 3$ in more than 50% of replicates, but the more parsimonious neighbourhood structure ($H = 2$) yields more efficient estimators *on average*, although only by a narrow margin and not uniformly for all the counties.

The summaries $f^{(\mathrm{AB})}$, $r^{(\mathrm{AB})}$ and $\#^{(\mathrm{AB})}$ can be adapted to reflect the greater importance of gains in efficiency for the less populous counties by associating them with unequal weights. At an extreme, we may focus on the counties up to a certain population size and ignore the rest. Although the gains or losses for the most populous counties are modest with all the methods (*vis-à-vis* direct estimation), some methods may be particularly effective for the least populous counties.

7.1. Conclusion

The method described and applied in this article combines design-based (direct) estimation and a distribution-free model that relates the degree of similarity of the target quantities θ_d to the distances of the corresponding districts. The consequence of this model is that the auxiliary information for a district is 'packaged' within the rings (sets of equidistant districts) around the target district. Any reference to a model can be completely dispensed with; for inference about a particular district, we regard the districts as more or less relevant depending on their distance from it.

Composition is a general principle, applicable whenever there are alternative estimators of a target. It requires no model and does not rely on any asymptotics. The combination of the estimators is *target-specific*; the coefficients depend on the (estimated) joint distribution of the basis estimators. Having to estimate the coefficients of the ideal composition is a drawback of the composition, comparable to the uncertainty associated with the estimation of the model parameters and with the validity of the model in an EB approach.

There are no constraints on how the distance is defined, although each value of the distance should occur for many pairs of districts, so that the covariances γ_h are estimated with high precision and most districts have several districts in their rings for each distance. In our application, we found that the estimation of even a single covariance, γ_1, when we distinguish only between neighbours and non-neighbours, introduces a lot of uncertainty in the estimation of the targets, and the estimation of the MSE is degraded a lot in comparison with the estimators that ignore the distance. Parsimony issues apply equally to the number of distinct distances H and to the choice of auxiliary variables to be used, as they do in model-based estimation. In applications not reported here, we found that setting $H = 2$ is sufficient and defining the distance $\xi = 1$ for geographical neighbours to be adequate.

Unlike ML, composite estimation involves no iterations, and so even more intensive simulations can be conducted with it. All the computing described in this article was conducted in R, R Development Core Team (2007), and the code, in the form of functions, can be obtained from the author on request.

Acknowledgements

Support by the Grant SEJ–2006–13537 from the Spanish Ministry of Science and Technology is acknowledged. Àlex Costa and Dolors Olivares from IDESCAT, Barcelona, assisted with background information; Anna Cuxart, Xavier Palacios and Joan Pallarés provided some useful comments and suggestions on an earlier version of the manuscript.

Appendix A. Derivation of the identity in (8)

We express $\theta_d^{(h)}$ in terms of the district-level quantities θ_d and expand the square:

$$\mathrm{E}_{\mathcal{D}}\left(\Delta^2\theta_d^{(h)}\right) = \frac{1}{N_d^{(h)2}}\sum_{d'\in\mathbf{d}_d^{(h)}} N_{d'}^2\,\mathrm{E}_{\mathcal{D}}\left\{(\theta_{d'}-\theta_d)^2\right\} + \frac{1}{N_d^{(h)2}}\sum_{d_1}\sum_{d_2} N_{d_1}N_{d_2}\,\mathrm{E}_{\mathcal{D}}\left\{\left(\theta_{d_1}-\theta_d\right)\left(\theta_{d_2}-\theta_d\right)\right\}, \tag{14}$$

where the double summation is over the pairs of districts $d_1 \neq d_2$, both of which belong to $\mathbf{d}_d^{(h)}$. Since $\xi(d, d') = h$ for every $d' \in \mathbf{d}_d^{(h)}$, the first summation is equal to

$$\sum_{d'\in\mathbf{d}_d^{(h)}} N_{d'}^2\,\mathrm{E}_{\mathcal{D}}\left\{(\theta_{d'}-\theta_d)^2\right\} = \sigma_h^2\sum_{d'\in\mathbf{d}_d^{(h)}} N_{d'}^2.$$

An expectation in the double summation in (14) is expressed in terms of the variances σ_h^2 and covariances γ_h as

$$\mathrm{E}_{\mathcal{D}}\left\{(\theta_{d_1}-\theta)(\theta_{d_2}-\theta) - (\theta_{d_1}-\theta)(\theta_d-\theta) - (\theta_{d_2}-\theta)(\theta_d-\theta) + (\theta_d-\theta)^2\right\} = \gamma_{\xi(d_1,d_2)} - 2\gamma_h + \sigma_0^2.$$

Please cite this article in press as: Longford, N.T., Small area estimation with spatial similarity. Computational Statistics and Data Analysis (2009), doi:10.1016/j.csda.2009.09.005

Table 4
Composite estimators. Definitions and notation.

Estimator	Notation	Description
Direct	$\hat{\theta}_d$	Based on the data from district d only
Composite (general)	$\tilde{\theta}_d$	Convex combination of basis estimators
Univariate composite		
– without spatial similarity	$\tilde{\theta}_d^{(1)}$	*U-Comp-1* — composition of the direct estimators of the target district $\hat{\theta}_d$ and of its complement $\hat{\theta}_{-d}$
– with neighbourhood similarity	$\tilde{\theta}_d^{(2)}$	*U-Comp-2* — composition of the direct estimators of the target district, its neighbourhood (1-ring) and the remainder of the country (2-ring)
– with spatial similarity (general)	$\tilde{\theta}_d^{(H)}$	*U-Comp-H* — composition of the direct estimators of the target district and of its h-rings, $1 \le h \le H$
Bivariate composite		
– without spatial similarity	$\tilde{\theta}_d^{(1)}$	*B-Comp-1* — bivariate composition of the direct estimator and auxiliary information for the target district, $\hat{\boldsymbol{\theta}}_d$, with vector $\hat{\boldsymbol{\theta}}_{-d}$ for the complement
– with neighbourhood similarity	$\tilde{\theta}_d^{(2)}$	*B-Comp-2* — bivariate composition of the vector $\hat{\boldsymbol{\theta}}_d$ with corresponding vector $\hat{\boldsymbol{\theta}}_d^{(1)}$ for the neighbours (1-ring) and $\hat{\boldsymbol{\theta}}_d^{(2)}$ for the 2-ring of district d
– with spatial similarity (general)	$\tilde{\theta}_d^{(H)}$	*B-Comp-H* — bivariate composition of the vector $\hat{\boldsymbol{\theta}}_d$ with the vectors $\hat{\boldsymbol{\theta}}_d^{(h)}$ for the h-rings of district d; $1 \le h \le H$
Multiariate composite (general)		Composition of $\hat{\boldsymbol{\theta}}_d$, containing multivariate auxiliary information, with its counterpart vectors $\hat{\boldsymbol{\theta}}_d^{(h)}$ for the h-rings of district d

After completing the double summation in (14) by the 'diagonal' contributions that correspond to $d_1 = d_2$, we obtain the identity

$$\begin{aligned}\mathrm{E}_{\mathcal{D}}\left(\Delta^2\theta_d^{(h)}\right) &= M_d^{(h)}\sigma_h^2 + \mathbf{r}_d^{(h)\top}\boldsymbol{\Gamma}_d^{(h)}\mathbf{r}_d^{(h)} - M_d^{(h)}\sigma_0^2 + \left(\sigma_0^2 - 2\gamma_h\right)\left(1 - M_d^{(h)}\right)\\ &= \mathbf{r}_d^{(h)\top}\boldsymbol{\Gamma}_d^{(h)}\mathbf{r}_d^{(h)} + M_d^{(h)}\left(\sigma_h^2 - 2\sigma_0^2 + 2\gamma_h\right) + \sigma_0^2 - 2\gamma_h\\ &= \mathbf{r}_d^{(h)\top}\boldsymbol{\Gamma}_d^{(h)}\mathbf{r}_d^{(h)} + \sigma_0^2 - 2\gamma_h,\end{aligned}$$

where $M_d^{(h)} = \mathbf{r}_d^{(h)\top}\mathbf{r}_d^{(h)}$, and $\boldsymbol{\Gamma}_d^{(h)}$ and $\mathbf{r}_d^{(h)}$ are as defined in (8).

Appendix B. An expression for var($\hat{\boldsymbol{\psi}}^\top\mathbf{G}\hat{\boldsymbol{\psi}}$)

The derivation follows in outline the proof of Theorem 1.8 in Seber (1977), in which a similar statement is proved. We assume that the direct estimators $\hat{\theta}_d$ are independent and that each has a symmetric sampling distribution.

Recall that $\boldsymbol{\psi} = (\theta_1, \ldots, \theta_D)^\top$ and $\hat{\boldsymbol{\psi}}$ is the vector of the corresponding direct estimators, so that $\mathrm{E}_{\mathcal{S}}\left(\hat{\boldsymbol{\psi}}\right) = \boldsymbol{\psi}$. Further, let $\mathbf{q}$ be the vector for which $\hat{\theta} = \mathbf{q}^\top\hat{\boldsymbol{\psi}}$. Both $\hat{\sigma}_{\mathrm{A}}^2$ and $\hat{\sigma}_{\mathrm{B}}^2$ have the form $\hat{\boldsymbol{\psi}}^\top\mathbf{G}\hat{\boldsymbol{\psi}} - e$, with $\mathbf{G} = (\mathbf{I}_D - \mathbf{1}\mathbf{q}^\top)^\top\mathbf{U}(\mathbf{I}_D - \mathbf{1}\mathbf{q}^\top)$. For $\hat{\sigma}_{\mathrm{A}}^2$, $\mathbf{U} = D^{-1}\mathbf{I}$, and for $\hat{\sigma}_{\mathrm{B}}^2$, $\mathbf{U} = \mathrm{diag}(\mathbf{r})$, where $\mathbf{r}$ is the vector of sampling fractions n_d/N_d. Let $\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\psi}} - \boldsymbol{\psi}$. Then

$$\begin{aligned}\mathrm{var}\left(\hat{\boldsymbol{\psi}}^\top\mathbf{G}\hat{\boldsymbol{\psi}}\right) &= \mathrm{var}\left(\hat{\boldsymbol{\gamma}}^\top\mathbf{G}\hat{\boldsymbol{\gamma}} + 2\boldsymbol{\psi}^\top\mathbf{G}\hat{\boldsymbol{\gamma}}\right)\\ &= \mathrm{var}\left(\hat{\boldsymbol{\gamma}}^\top\mathbf{G}\hat{\boldsymbol{\gamma}}\right) + 4\boldsymbol{\psi}^\top\mathbf{G}\mathbf{V}\mathbf{G}\boldsymbol{\psi} + 4\mathrm{cov}\left(\hat{\boldsymbol{\gamma}}^\top\mathbf{G}\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\gamma}}\right)\mathbf{G}\boldsymbol{\psi}. \qquad (15)\end{aligned}$$

The expansion of the covariance to terms $\mathrm{cov}(G_{ij}\hat{\gamma}_i\hat{\gamma}_j, \hat{\gamma}_k)$ comprises the expectations of products of powers of $\hat{\gamma}_d$, at least one of which has an odd exponent. Owing to the symmetry of the underlying distribution, each such term vanishes, and therefore so does the covariance in (15).

Further, $\mathrm{var}(\hat{\boldsymbol{\gamma}}^\top\mathbf{G}\hat{\boldsymbol{\gamma}}) = \mathrm{E}(\hat{\boldsymbol{\gamma}}^\top\mathbf{G}\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}^\top\mathbf{G}\hat{\boldsymbol{\gamma}}) - \{\mathrm{tr}(\mathbf{G}\mathbf{V})\}^2$. In the expansion of the expectation to a four-way summation, only the terms that involve γ_d^4 and $\gamma_{d_1}^2\gamma_{d_2}^2$ are non-zero, and so

$$\mathrm{E}\left(\hat{\boldsymbol{\gamma}}^\top\mathbf{G}\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}^\top\mathbf{G}\hat{\boldsymbol{\gamma}}\right) = \sum_{d=1}^{D} G_{dd}^2\,\mathrm{E}\left(\hat{\gamma}_d^4\right) + \sum_{d_1}\sum_{d_2} G_{d_1d_1}G_{d_2d_2}\,v_{d_1}\,v_{d_2} + 2\sum_{d_1}\sum_{d_2} G_{d_1d_2}^2\,v_{d_1}\,v_{d_2},$$

where each double summation is over the pairs of distinct districts ($d_1 \ne d_2$). They correspond to the three kinds of pairwise agreements among four subscripts. Adding to these summations the 'diagonal' terms, which correspond to $d_1 = d_2$ and are equal to $\sum_d G_{dd}^2\,v_d^2$ in each instance, yields the expression

$$\mathrm{E}\left(\hat{\boldsymbol{\gamma}}^\top\mathbf{G}\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}^\top\mathbf{G}\hat{\boldsymbol{\gamma}}\right) = \sum_{d=1}^{D}(\kappa_d - 3)\,G_{dd}^2\,v_d^4 + \{\mathrm{tr}\,(\mathbf{G}\mathbf{V})\}^2 + 2\mathbf{v}^\top\mathbf{G}^2\mathbf{v},$$

Please cite this article in press as: Longford, N.T., Small area estimation with spatial similarity. Computational Statistics and Data Analysis (2009), doi:10.1016/j.csda.2009.09.005

where $\mathbf{v} = (v_1, \ldots v_D)^\top$ is the diagonal of $\mathbf{V}$. Therefore

$$\text{var}\left(\hat{\boldsymbol{\psi}}^\top \mathbf{G}\hat{\boldsymbol{\psi}}\right) = \sum_{d=1}^{D} (\kappa_d - 3)\, G_{dd}\, v_d^4 + 2\mathbf{v}^\top \mathbf{G}^2\mathbf{v} + 4\boldsymbol{\psi}^\top \mathbf{G}\mathbf{V}\mathbf{G}\boldsymbol{\psi}. \tag{16}$$

The summation that involves the kurtoses vanishes when the estimators $\hat{\theta}_d$ are normally distributed. We also require the identity

$$\text{cov}\left(\hat{\boldsymbol{\psi}}^\top \mathbf{G}_\text{A}\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\psi}}^\top \mathbf{G}_\text{B}\hat{\boldsymbol{\psi}}\right) = \sum_{d=1}^{D} (\kappa_d - 3)\, G_{\text{A},dd} G_{\text{B},dd}\, v_d^4 + 2\mathbf{v}^\top \mathbf{G}_\text{A}\mathbf{G}_\text{B}\mathbf{v} + 4\boldsymbol{\psi}^\top \mathbf{G}_\text{A}\mathbf{V}\mathbf{G}_\text{B}\boldsymbol{\psi}.$$

It is derived directly by substituting (16) in the identity

$$\text{cov}\,(Y_\text{A}, Y_\text{B}) = \frac{1}{4}\left\{\text{var}\,(Y_\text{A} + Y_\text{B}) - \text{var}\,(Y_\text{A} - Y_\text{B})\right\},$$

where $Y_\text{C} = \hat{\boldsymbol{\psi}}^\top \mathbf{G}_\text{C}\hat{\boldsymbol{\psi}}$ and C = A, B.

References

Congdon, P., 2004. Modelling trends and inequality in small-area mortality. Journal of Applied Statistics 31, 603–622.
Efron, B., Morris, C.N., 1972. Limiting the risk of Bayes and empirical Bayes estimators—part II: The empirical Bayes case. Journal of the American Statistical Association 67, 130–139.
Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap. Chapman and Hall, London.
Elliott, P., Wakefield, J., 2001. Disease clusters: Should they be investigated, and, if so, when and how?. Journal of the Royal Statistical Society Series A 164, 3–12.
Fay, R.E., Herriot, R.A., 1979. Estimates of income for small places: An application of James-Stein procedures to census data. Journal of the American Statistical Association 74, 269–277.
Goldstein, H., 2002. Multilevel Statistical Models, 3rd ed. Edward Arnold, London.
James, W., Stein, C., 1961. Estimation with quadratic loss. In: Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 361–379.
Kang, E.L., Liu, D., Cressie, N., 2009. Statistical analysis of small-area data based on independence, spatial, non-hierarchical, and hierarchical models. Computational Statistics and Data Analysis 53, 3016–3032.
Longford, N.T., 1993. Random Coefficient Models. Oxford University Press, Oxford, UK.
Longford, N.T., 1999. Multivariate shrinkage estimation of small area means and proportions. Journal of the Royal Statistical Society Series A 162, 227–245.
Longford, N.T., 2000. On estimating standard errors in multilevel analysis. The Statistician 49, 389–398.
Longford, N.T., 2004. Missing data and small area estimation in the UK Labour Force Survey. Journal of the Royal Statistical Society Series A 167, 341–373.
Longford, N.T., 2005. Missing Data and Small-Area Estimation. Analytical Equipment for the Survey Statistician. Springer-Verlag, New York.
Longford, N.T., 2007. On standard errors of model-based small-area estimators. Survey Methodology 33, 69–79.
Longford, N.T., 2008. Small-area estimation with spatial similarity. Working Paper No. 1105, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Barcelona, Spain.
McCullagh, P., 2008. Sampling bias in logistic regression. Journal of the Royal Statistical Society Series B 70, 643–677.
Nelder, J.A., Lee, Y., Pawitan, Y., 2006. Generalized Linear Models with Random Effects: A Unified Approach via h-Likelihood. Chapman and Hall, London.
Pfeffermann, D., Tiller, R., 2006. State-space modelling with correlated measurement errors with application to small area estimation under benchmark constraints. Journal of the American Statistical Association 101, 1387–1397.
Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-plus. Springer-Verlag, New York.
Potthoff, R.F., Woodbury, M.A., Manton, K.G., 1992. "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. Journal of the American Statistical Association 87, 383–396.
Prasad, N.G.N., Rao, J.N.K., 1990. The estimation of the mean squared error of small area estimators. Journal of the American Statistical Association 85, 163–171.
Rao, J.N.K., 2003. Small Area Estimation. Wiley, New York.
R Development Core Team,, 2007. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.
Särndal, C.-E., Swensson, B., Wretman, J., 1992. Model Assisted Survey Sampling. Springer-Verlag, New York.
Seber, G.A.F., 1977. Linear Regression Analysis. Wiley, New York.
Temiyasathit, C., Kim, S.B., Park, S.-K., 2009. Spatial prediction of ozone concentration profiles. Computational Statistics and Data Analysis 53, 3892–3906.

Please cite this article in press as: Longford, N.T., Small area estimation with spatial similarity. Computational Statistics and Data Analysis (2009), doi:10.1016/j.csda.2009.09.005