An assessment of empirical Bayes and composite estimators for small areas

Nicholas T. Longford, SNTL, Leicester, England

Summary

We compare a set of empirical Bayes and composite estimators of the population means of the districts (small areas) of a country, and show that the natural modelling strategy of searching for a well fitting empirical Bayes model and using it for estimation of the area-level means can be inefficient.

Key words: Composite estimator, empirical Bayes models, mean squared error, small-area estimation.

JEL Classification: C15 — Econometric and statistical methods: Simulation methods

Address for correspondence: N. T. Longford, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25–27, Barcelona, Spain. Email: NTL@SNTL.co.uk .

1 Introduction

Research in small-area estimation has in the recent years received a strong impetus from several communities involved in the analysis of large-scale national surveys. Inferences that were in the past made only at the national and regional levels are now sought also for smaller administrative units, such as counties or districts (small areas). Much of small-area estimation uses a model-based approach in which a suitable model is fitted to the survey data and district-level estimates are derived from the model fit. Empirical Bayes (EB) models are particularly well suited for this purpose because the estimates they yield borrow strength across the districts (exploit their similarity) and adjust for differences in the background variables recorded by the survey. In this paper, we consider an outcome variable with normal conditional distribution, given the values of covariates and district-level deviations. The outcome and covariates are related by the EB model

$$y_{id} \;=\; \mathbf{x}_{id}\boldsymbol{\beta} + \delta_d + \varepsilon_{id}\,, \tag{1}$$

where ε_{id}, $i = 1, \ldots, n_d$, and δ_d, $d = 1, \ldots, D$, are mutually independent random samples from centered normal distributions with respective variances σ^2_{W} and σ^2_{B}.

The model fit comprises estimates $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2_{\mathrm{W}}$ and $\hat{\sigma}^2_{\mathrm{B}}$ and estimated conditional expectations $\hat{\delta}_d \;=\; \mathrm{E}\left(\delta_d \,|\, \boldsymbol{\beta} = \hat{\beta}, \sigma^2_{\mathrm{W}} = \hat{\sigma}^2_{\mathrm{W}}, \sigma^2_{\mathrm{B}} = \hat{\sigma}^2_{\mathrm{B}}\right)$, which can be regarded as district-level residuals. The district-level population mean of Y, denoted by $\bar{Y}_d$, is estimated by $\bar{\mathbf{x}}_d\hat{\boldsymbol{\beta}} + \hat{\delta}_d$, where $\bar{\mathbf{x}}_d$ is the vector of sample means of the covariates. When the population mean of a covariate is available, it is substituted in $\bar{\mathbf{x}}_d$ for the corresponding sample mean. We refer to the combination of a model (set of covariates) and the available district-level population information (sets of population means) as a *setting*.

A typical modelling strategy sets out with a search for a well fitting model, presuming that models that can be regarded as valid are associated with efficient inference. We show by example that this strategy is not always appropriate, and contrast it with composite estimation which bypasses fitting models such as (1). The example uses an artificial population, the construction of which is loosely based on the labour force of Slovakia; details of this population are given in Section 1.1. Three variables are considered, a continuous outcome variable Y (recent monthly log-wage), a similar covariate X (log-wage a year ago), and a dichotomous variable Z with the within-district probabilities in the range $0.50 - 0.60$, based on gender. All three variables are defined for members of the labour force, the elementary units in the survey. The survey has a stratified sampling design with sample size $n = 4000$. The 79 districts of Slovakia (*okresy*) are the strata and

simple random sampling with sample size fixed and proportional to the population size is applied, independently, in each district. The sizes of the districts vary a great deal, so that their sample sizes are in the range 9 – 122.

Seven estimators are considered, corresponding to models with covariates X and Z, or only X, and availability of the values of the district-level population means $\bar{X}_d$ and $\bar{Z}_d$ of X and Z, respectively, from an external source. For the model with X and Z as the covariates, we consider settings with neither sets of these means available, available $\bar{Z}_d$ but not $\bar{X}_d$, and both $\bar{X}_d$ and $\bar{Z}_d$ available. For the model with X, the values of $\bar{X}_d$ may all be either available or not. We denote these settings by indicating the covariates in the model and the available population information. Thus $(X, Z \mid \bar{X}_d)$ stands for the model with X and Z, with $\bar{X}_d$ available, but $\bar{Z}_d$ not. The district-level population means $\bar{X}_d$ and $\bar{Z}_d$ may also be used as covariates. The vertical bar is omitted when no district-level population information is available beyond what is included in the model. For instance, (X, Z) stands for the model with X and Z as the covariates, but neither $\bar{X}_d$ nor $\bar{Z}_d$ available. In the setting $(\bar{X}_d)$, $\bar{Z}_d$ is not available, but $\bar{X}_d$ is, and is used as the only covariate.

Counterparts of these model-based estimators can be defined in composite estimation. For example, the composite estimator that corresponds to $(X, Z \mid \bar{Z}_d)$ seeks the multivariate convex combination of the district-level sample means of X, Y and Z and of $\bar{Z}_d$ with the vector of the corresponding national means: the sample means of X, Y and Z and the (national) population mean of Z. Details are given in the Appendix.

The properties of these two sets of estimators, composite and EB, are established empirically, by replications of the sampling and estimation processes. We find that even though Z is an important predictor of Y, using it for estimation of the district-level means $\bar{Y}_d$ is counterproductive when $\bar{Z}_d$ are not available. That is, the EB estimators of the district-level means $\bar{Y}_d$ based on (X, Z) are inferior to those based on (X). The contradiction is absent in composite estimators which, on average, are more efficient than their model based counterparts, when the values of $\bar{X}_d$ are not available. When X is included in the model and the district-level population means $\bar{X}_d$ are available, the EB estimators based on the settings $(X, Z \mid X_d, Z_d)$ and $(X \mid X_d)$ are on average more efficient than their composition counterparts. The results are discussed in more detail in Section 2.

1.1 The simulated population

Slovakia is one of the two countries formed when Czechoslovakia split in 1993. At the census in 2001, its population was 5.379 million, 2.666 million (50%) of them economically active, and of these 48% were women. Administratively, Slovakia is divided into eight counties (*kraje*), and these are further divided into between 7 to 13 districts. There are 79 districts in total, nine of them are parts of the two largest cities, Bratislava and Košice.

The population sizes of the districts are in the range 13 000 – 150 000; their mean is 68 100 and median 61 900. For $\bar{X}_d$ we use the logarithm of the mean monthly wage in 2005. This is available from the official website of the Slovak Statistical Office

(`http://www.statistics.sk/webdata/ks/ksbrat/mesacmzda.htm`)

only for the counties, so we generate a single set of values of $\bar{X}_d$ by adding to them deviations drawn at random from $\mathcal{N}(0, 0.05^2)$. This corresponds to about ±5% on the linear scale, our guess of the variation of the district-level means within the counties. The county that contains the capital Bratislava had by far the highest mean income, Sk23 200 (*Slovenská koruna*) in 2005, 34% higher than the national average of Sk17 200. (In 2005, £1 was equivalent to about Sk55.) The means for all the other counties, in the range Sk13 200 – 16 800, are below the national average. On the log-scale, Sk23 200 corresponds to 10.052 and Sk13 200 to 9.488. The within-district standard deviations of log-wage were generated as $0.35 + \mathcal{N}\left(0, 0.015^2\right) + \frac{1}{6} S_d \cdot 10^{-6}$, where S_d is the population size of the district. The standard deviations are in the range 0.33 – 0.39 and their correlation with the population size is 0.40 — more populous districts tend to have more heterogeneous log-wages. The mean log-wage is correlated with the population size more weakly; their correlation is 0.08. The percentage of men in the labour force (Z_d) was generated as $7\mathcal{U}^{1.25} - 0.5^{1.25} + 56.0 - S_d/30\,000$, where $\mathcal{U}$ denotes the uniform distribution on $(0, 1)$. The percentages are in the range 51.7 – 61.1% and their correlation with the district-level population sizes is 0.50. The model deviations δ_d were generated as a random sample from $\mathcal{N}(0, 0.0025)$, so that the districts differ from the prediction $\bar{\mathbf{x}}_d \beta$ based on (1) by about 5% on average. The values of $\bar{X}_d$ and $\bar{Z}_d$ as well as the district-level deviations δ_d in (1) are fixed in the simulations — each replication is based on the same country with the same division into districts and with members of the labour force in the same district and with the same income. The sampling process is the only source of variation among the replicate samples. The values of the outcome variable are generated according to the model in (1) with intercept 2.0, slope on X equal to 0.8, and difference between men and women set to 0.25. For orientation, the values of $\bar{X}_d$, $\bar{Z}_d$ and δ_d and other information are plotted in Figure 1 and the exact values are listed in Table 1. The districts are not identified because the population means generated for them may differ substantially from their genuine population counterparts. The subject-level (residual) variance in (1) is set to 0.0625.

In a replication, independent random samples with set sample sizes are drawn from the joint distribution of X and Z in each district. As the sampling fraction is smaller than 0.001, the sampling can for all purposes be regarded as from an infinite population. The outcomes Y are generated according to the model in (1), but with fixed values of δ_d. The seven sets of estimates (79×7 matrices) are then evaluated for the EB method and composition. To streamline the calculations, we evaluate their deviations from the targets

Figure 1: Pairwise plots of the district-level variables: population size of the district, in thousands; population mean of X (past log-wage); within-district standard deviation of X; population mean of Y (recent log-wage); the model deviation δ_d in (1); percentage of men ($Z = 1$) in the labour force.

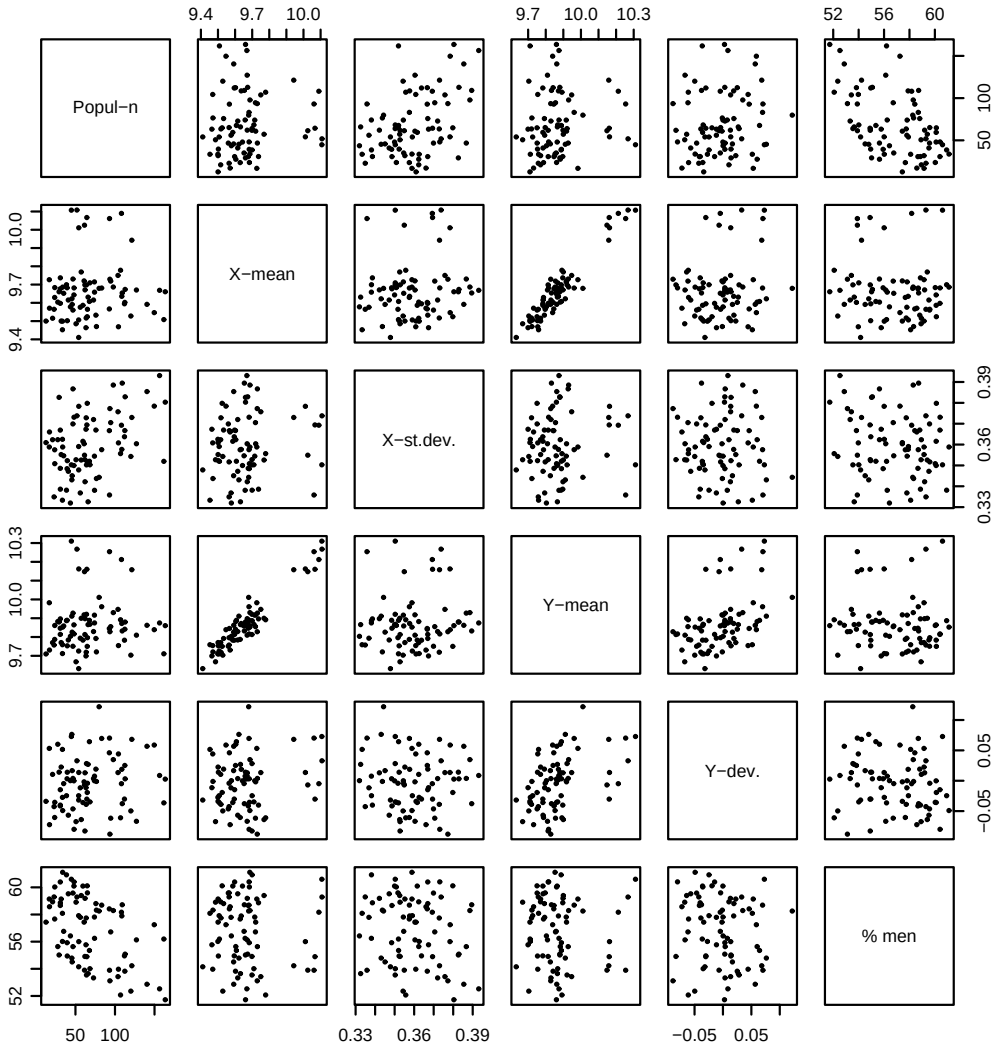


Table 1: District-level information used in the simulations. See caption of Figure 1 for the variable labels.

No.	Population	$\bar{X}_d$	$\sqrt{\text{var}\,(X_{id} \mid d)}$	$\bar{Y}_d$	δ_d	$100\bar{Z}_d$ (%)	County
1	44 798	10.107	0.350	10.310	0.073	60.599	A
2	108 139	10.090	0.369	10.212	-0.005	58.169	A
3	61 418	10.025	0.355	10.148	-0.007	53.898	A
4	93 058	10.061	0.336	10.254	0.070	53.894	A
5	121 259	9.943	0.373	10.158	0.068	54.225	A
6	64 354	10.067	0.369	10.160	-0.030	54.893	A
7	54 164	10.011	0.378	10.162	0.014	56.002	A
8	51 825	10.108	0.374	10.268	0.033	59.289	A
9	112 384	9.602	0.367	9.842	0.025	53.821	B
10	94 533	9.719	0.371	9.855	-0.062	56.718	B
11	45 351	9.622	0.343	9.911	0.076	54.791	B
12	63 928	9.698	0.350	9.921	0.028	53.544	B
13	60 891	9.720	0.360	9.918	-0.002	57.437	B
14	46 791	9.730	0.387	9.927	0.004	55.635	B
15	127 125	9.670	0.360	9.810	-0.067	56.130	B
16	38 640	9.699	0.338	9.887	-0.025	60.924	C
17	62 042	9.671	0.342	9.898	0.013	59.190	C
18	29 243	9.660	0.383	9.880	0.014	55.017	C
19	63 530	9.517	0.373	9.719	-0.039	57.742	C
20	48 005	9.643	0.350	9.865	0.000	60.095	C
21	65 150	9.565	0.352	9.797	-0.004	59.595	C
22	140 444	9.592	0.385	9.862	0.057	52.864	C
23	45 761	9.591	0.345	9.826	0.017	54.508	C
24	112 767	9.673	0.383	9.880	0.003	55.080	C
25	108 556	9.636	0.358	9.872	0.019	57.893	D
26	120 021	9.528	0.354	9.703	-0.050	52.354	D
27	163 540	9.661	0.380	9.861	0.003	51.721	D
28	149 594	9.546	0.378	9.839	0.059	57.247	D
29	54 000	9.506	0.359	9.725	-0.025	57.916	D
30	74 089	9.658	0.338	9.875	0.008	56.246	D
31	43 622	9.578	0.332	9.804	0.000	56.422	E
32	30 788	9.597	0.367	9.814	-0.015	60.396	E
33	92 843	9.567	0.367	9.845	0.046	58.416	E
34	39 364	9.682	0.348	9.889	0.003	55.922	E
35	33 778	9.686	0.359	9.853	-0.049	61.110	E
36	73 984	9.715	0.354	9.916	0.000	57.813	E
37	97 813	9.687	0.389	9.930	0.035	58.293	E
38	56 053	9.576	0.337	9.794	-0.012	57.791	E
39	59 420	9.586	0.362	9.783	-0.036	60.096	E
40	16 866	9.727	0.359	9.982	0.053	59.142	E
41	35 062	9.652	0.355	9.859	-0.005	56.759	E
42	156 361	9.669	0.393	9.875	0.009	52.523	E
43	111 984	9.595	0.364	9.783	-0.030	54.952	F
44	17 151	9.571	0.366	9.731	-0.072	58.574	F
45	65 909	9.631	0.333	9.866	0.027	53.647	F
46	33 514	9.499	0.363	9.712	-0.027	56.014	F
47	22 885	9.565	0.335	9.757	-0.043	59.294	F
48	72 837	9.615	0.372	9.829	0.004	53.328	F
49	23 666	9.629	0.357	9.787	-0.061	57.669	F
50	40 918	9.615	0.347	9.784	-0.057	59.527	F
51	83 124	9.681	0.358	9.961	0.070	58.690	F
52	46 741	9.560	0.370	9.751	-0.046	59.579	F
53	67 633	9.607	0.352	9.889	0.064	55.351	F
54	27 634	9.604	0.362	9.882	0.060	55.639	F
55	48 125	9.682	0.373	9.816	-0.080	60.406	F
56	75 793	9.533	0.367	9.792	0.019	58.806	G
57	64 845	9.524	0.380	9.750	-0.019	60.113	G
58	63 231	9.464	0.367	9.699	-0.011	55.770	G
59	31 880	9.500	0.355	9.771	0.022	59.141	G
60	12 668	9.500	0.361	9.710	-0.034	57.436	G
61	104 348	9.469	0.361	9.754	0.044	53.970	G
62	161 782	9.508	0.352	9.711	-0.036	56.193	G
63	54 067	9.411	0.348	9.632	-0.032	54.150	G
64	39 633	9.502	0.352	9.753	0.014	54.941	G
65	50 684	9.485	0.353	9.668	-0.067	58.907	G
66	21 027	9.516	0.362	9.770	0.010	58.934	G
67	33 506	9.453	0.333	9.759	0.052	58.089	G
68	76 504	9.470	0.344	9.721	-0.001	58.323	G
69	30 841	9.737	0.339	9.896	-0.040	58.863	H
70	68 262	9.710	0.350	9.868	-0.035	54.008	H
71	79 850	9.679	0.344	10.011	0.122	58.255	H
72	30 745	9.677	0.349	9.847	-0.041	58.663	H
73	57 236	9.769	0.353	9.900	-0.064	59.399	H
74	106 999	9.778	0.356	9.892	-0.061	52.071	H
75	109 121	9.654	0.389	9.832	-0.038	58.719	H
76	61 887	9.698	0.353	9.813	-0.083	54.910	H
77	23 776	9.683	0.351	9.861	-0.036	60.028	H
78	93 516	9.729	0.377	9.828	-0.088	53.115	H
79	103 779	9.750	0.376	9.947	0.013	53.424	H

$\bar{Y}_d$ and accumulate their totals and totals of squares, to summarise the corresponding estimators by their empirical bias and root mean squared error (the square root of the mean squared error, rMSE).

2 Results

A simulation comprises 1000 replications, and its results are summarised by two $2 \times 7 \times 79$ arrays, one of biases and one of rMSEs, for the two methods (EB and composition), seven settings and 79 districts. We use the symbols `A`, `B`, `C`, `D`, `a`, `c` and `d` for the distinct settings and their estimators. In settings `A`, `B` and `C`, X and Z are the covariates and the district-level population means available are neither, only $\bar{Z}_d$, and for $(\bar{X}_d, \bar{Z}_d)$, respectively. The settings `a` and `c` are formed from `A` and `C`, respectively, by omitting Z and $\bar{Z}_d$. The settings `D` and `d` use only district-level covariates $\bar{X}_d$ and $\bar{Z}_d$.

Figure 2 displays the rMSEs, in separate panels for composite and EB estimation. The districts are presented in the ascending order of their sample sizes which are indicated in the right-hand margin of each panel. In the left-hand margins, the estimators with the smallest and largest rMSEs are indicated for each district (heading `W/L`, which stands for the 'winner' and 'loser'). The ragged solid line is drawn in both panels at $\sigma^2_{\mathrm{W}}/n_d$. Its sole purpose is to aid comparisons across the two panels of the diagram.

For the EB estimators, $(X, Z \,|\, \bar{X}_d, \bar{Z}_d)$ is the winner in 54 districts, (X) in 14 districts and $(\bar{X}_d)$ in the remaining 11 districts. The estimator (X, Z) is the loser in 71 districts. It is less efficient on average than estimator (X), so including Z in the model is detrimental when $\bar{Z}_d$ is not available. The model with covariates X and Z is valid (used for data generation), but the estimators of $\bar{Y}_d$ based on it are the losers for most districts when $\bar{X}_d$ and $\bar{Z}_d$ are not available. For the composite estimators, the picture is much less clear-cut; $(X \,|\, \bar{X}_d)$ is the winner in 34 districts and $(X, \bar{Z}_d \,|\, \bar{X}_d)$, $(\bar{X}_d)$ and $(X, Z \,|\, \bar{X}_d, \bar{Z}_d)$ in 13, 15 and 17 districts, respectively. The most frequent loser is the estimator $(X, Z \,|\, \bar{Z}_d)$, for 42 districts, but all the other estimators, except for $(X \,|\, \bar{X}_d)$, are losers, in 2 – 14 districts.

The array of results can be further summarised by the 2×7 table of average rMSEs, given in Table 2. The corresponding table of biases is of no importance because its entry vanishes for every estimator. The table confirms that using Z or $\bar{Z}_d$ is detrimental (on average) for EB estimators, whereas in composite estimation they make little difference; compare columns `A`, or `B`, with `a` and `D` with `d`. In fact, the rMSEs of the pairs of composite estimators `B` and `a` are very similar for all the districts; their differences are in the range $-0.0010 - 0.0013$. The differences of the pairs of composite estimators `D` and `d` are in the range $-0.0041 - 0.0018$. (There is much more overprinting in Figure 2 in the panel for composite estimators than for EB estimators.)

The bottom line of the table counts for each setting the number of districts for which the composite estimator is superior to its EB counterpart. It shows that the composite

Figure 2: The rMSEs of the composite estimators and EB estimators of the district-level means. The symbols are for the following settings: `A` — (X, Z), `B` — $(X, Z \mid \bar{Z}_d)$, `C` — $(X, Z \mid \bar{X}_d, \bar{Z}_d)$, `D` — $(\bar{X}_d, \bar{Z}_d)$, `a` — (X) and `c` — $(X \mid \bar{X}_d)$ and `d` — $(\bar{X}_d)$.

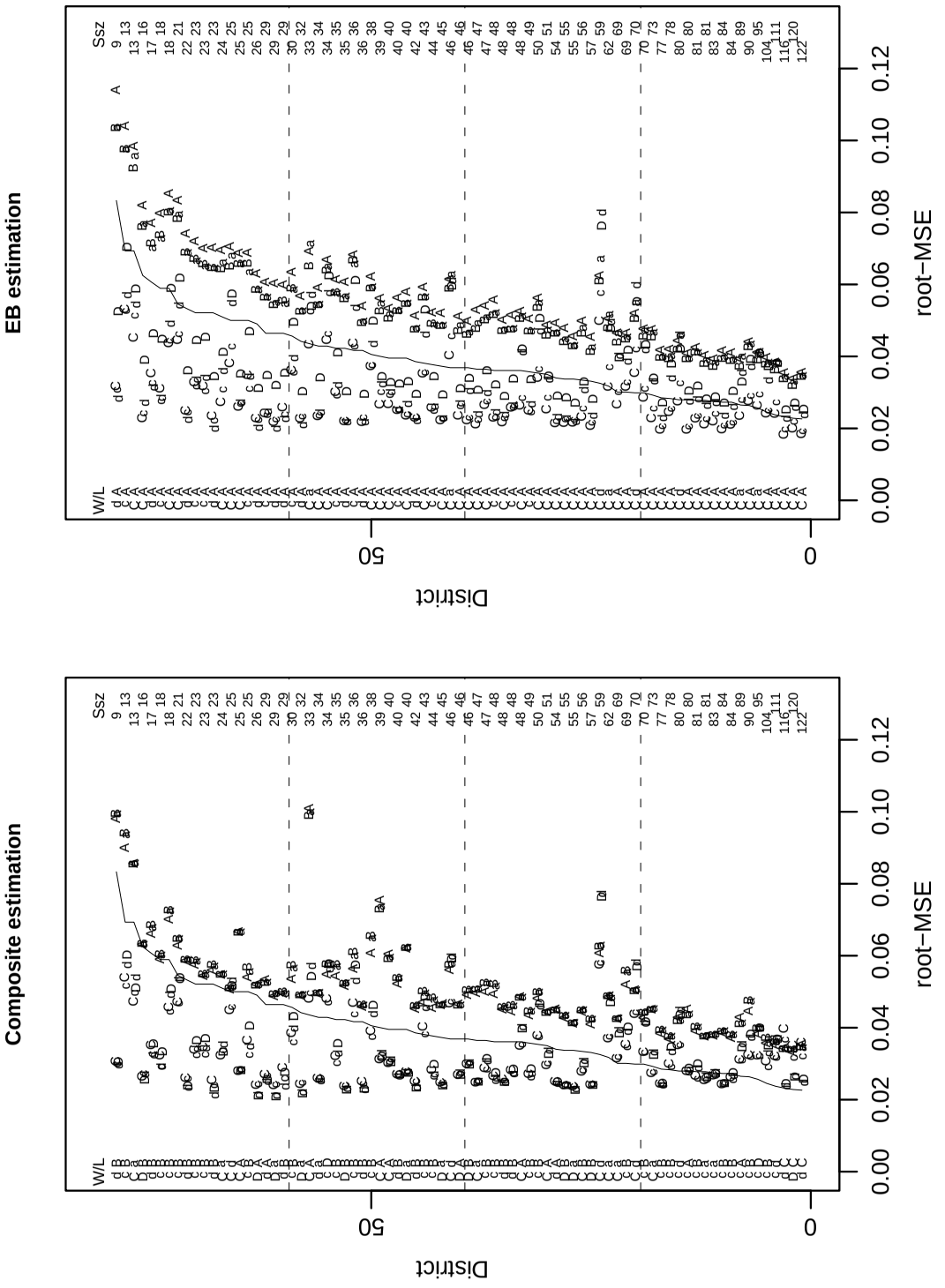


Table 2: The average rMSEs of the composite and empirical Bayes estimators and the numbers of districts for which composite estimator has smaller rMSE.

	Setting						
	A	B	C	D	a	c	d
	(X, Z)	$(X, Z \mid \bar{Z}_d)$	$(X, Z \mid \bar{X}_d, \bar{Z}_d)$	$(\bar{X}_d, \bar{Z}_d)$	(X)	$(X \mid \bar{X}_d)$	$(\bar{X}_d)$
Composite	0.0513	0.0523	0.0323	0.0346	0.0521	0.0316	0.0340
Emp. Bayes	0.0570	0.0536	0.0280	0.0394	0.0544	0.0292	0.0337
Comp. pref.	71	56	2	73	64	5	22

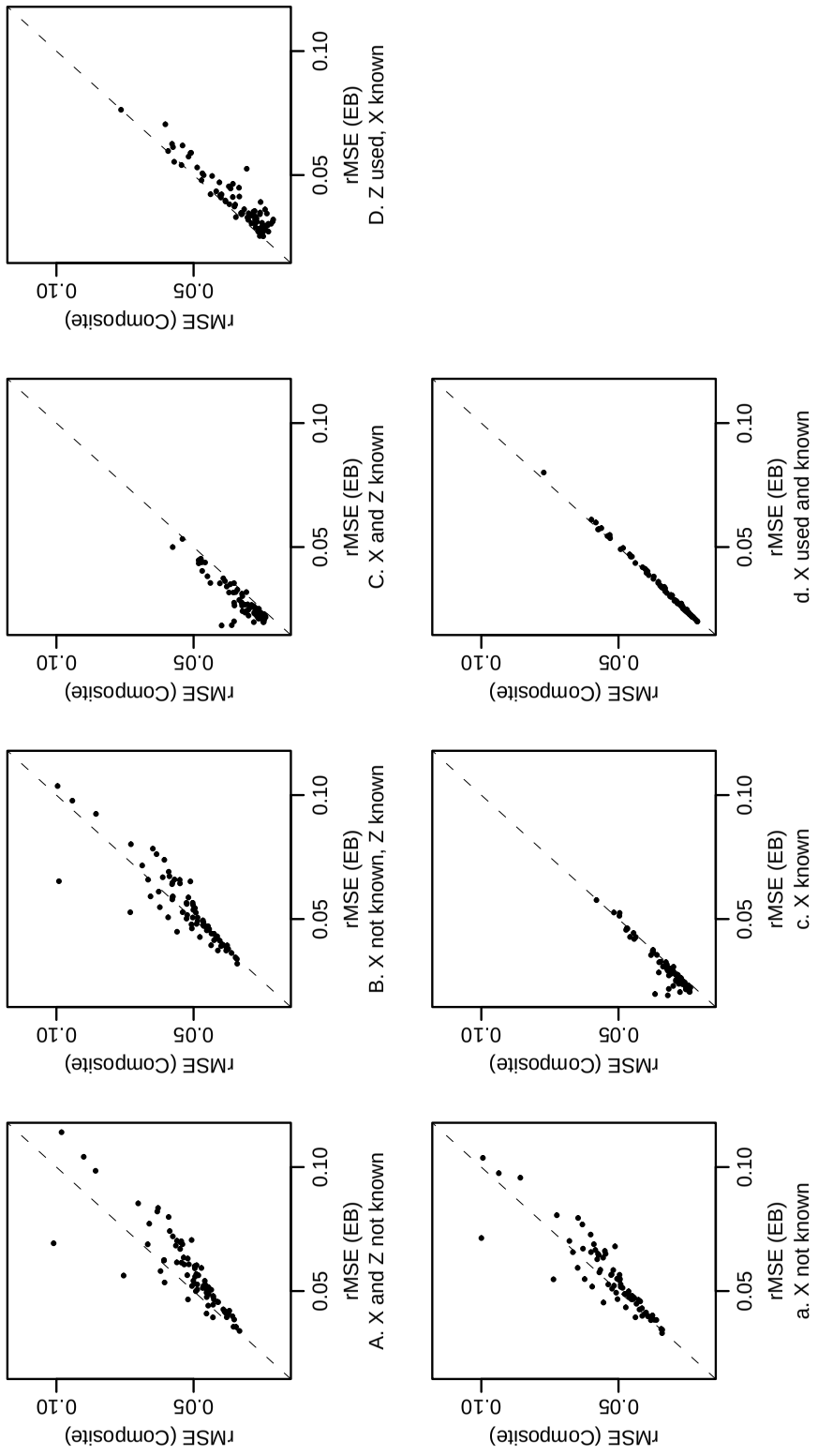
estimator is superior in a majority of districts for settings A, B, D and a, when the values of either X or $\bar{X}_d$ are not available. EB estimator is superior in a majority of districts in settings C and c, when both X and $\bar{X}_d$ are used. In setting d, the differences in rMSE are very small for all the districts. Figure 3 compares the pairs of estimators directly.

The MSE comprises two components: squared bias and variance. For each estimator, we can summarise these components by the ratio of squared bias and MSE, expressed as a percentage. In general, EB estimators have smaller biases, but their contributions to MSE are substantial for both sets of estimators. For example, for setting A, the average percentage contribution is 37.5% for EB and 59.8% for composite estimators. In contrast, for settings D (75%) and c (42%), they differ only slightly. For settings C, c and D, the contribution of the bias to MSE is greater for the composite estimators for the majority of districts (55, 69, and 68, respectively).

The comparisons of the estimators can be made more relevant when related to the available information. For example, when no district-level population information is available, only settings A, (X, Z), and a, (X), are relevant. The composite estimators are then preferred because they are more efficient for most of the districts for both settings A and a. When the values of $\bar{Z}_d$ are available setting B is relevant, although we may use A and a also, making no use of $\bar{Z}_d$. On average, setting A is most efficient for composite estimation, and setting B is most efficient for EB; the former is more efficient on average.

For composition, A is only slightly more efficient on average than a, even though it is more efficient for 69 districts (87%). Composition makes some use of the covariate Z. In contrast, for EB, A is less efficient than a for 71 districts; using the covariate Z is counterproductive. No model comparison would suggest this; the relevant likelihood ratio test statistics exceed 100.0 for all replicate datasets in the simulations (null distribution χ^2 with one degree of freedom). This shows that model selection followed by small-area estimation is problematic. On reflection, the source of this problem is obvious. Although the differences between men and women are substantial, their proportions within the districts vary relatively little, so that $\bar{z}_d \hat{\beta}_z$ contributes to the estimator of $\bar{Y}_d$ with a

Figure 3: The rMSEs of the composite and EB estimators of the district-level means for the seven settings `A`–`D` and `a`, `c`, and `d`.



lot of noise because the proportion $\bar{Z}_d$ is estimated by the sample proportion $\bar{z}_d$ with little precision. The problem is resolved when the values of $\bar{Z}_d$ are available, but the improvement is only slight. In fact, with composition, the average root-MSE for setting `B` is slightly higher than for `a`. For composition, $\bar{Z}_d$ is redundant. With composition, we may use the setting $(X, \bar{Z}_d)$, combining the sample means of X and Y and the district-level population proportions of Z. This setting is more efficient than `A`, but only slightly (average rMSE equal to 0.0502), but it does not have a natural EB counterpart.

When $\bar{X}_d$ is available the district-level means of Y are estimated with much greater precision by both EB and composition. EB is more efficient in the most elaborate setting `C` (by about 15% on average) and in setting `c` in which Z or Z_d are not used. For setting `d`, the two sets of estimators differ only slightly, and for setting `D`, composition is more efficient for most districts. We conclude that EB estimators are superior when the model in (1) includes the dominant predictor X and the district-level population means $\bar{X}_d$ are available. Without either of them, composite estimators are more efficient on average.

2.1 Strengths and weaknesses of the methods

Empirical Bayes estimators assume that the model in (1) is valid, that is, it includes the appropriate covariates, the random terms are homoscedastic and normally distributed, and the covariance structure is correctly specified. In (1), the covariance structure is compound symmetry, given by $\text{var}(\mathbf{y}_d) = \sigma^2_{\text{W}}\mathbf{I} + \sigma^2_{\text{B}}\mathbf{J}$, where $\mathbf{y}_d$ is the vector of outcomes for district d and $\mathbf{I}$ and $\mathbf{J}$ are the identity matrix and the matrix of ones of dimensions implied by the context (n_d). We constructed the population so that these assumptions are satisfied. Additionally, regression assumes that the values of the covariates are fixed. This assumption, often ignored, is not satisfied because each replication yields a different set of values of X and Z. Longford (2006a and b) shows that this assumption is innocuous for estimation for the means $\bar{Y}_d$, but not for estimation of the sampling variances or MSEs of the estimators by EB method or composition. According to EB theory, rMSEs are decreasing functions of the sample size, but Figure 2 clearly contradicts this for all settings.

In composite estimation, we assume that the vector $(\bar{X}_d, \bar{Y}_d, \bar{Z}_d)$ has a trivariate (district-level) distribution with a finite variance matrix. For a finite set of districts d, this assumption is vacuous because the distribution is not specified (e.g., by type and parameter space), so we can regard it as satisfied. The univariate distributions of X and Z, displayed in Figure 4, may raise some concern, because they are distinctly asymmetric; the mean log-income in one of the counties, which includes the capital, is much higher than in the rest of the country. Composite estimation has no distributional assumptions, although a variance matrix is in general regarded as an appropriate description of the dispersion for symmetric (and unimodal) distributions.

Figure 4: Stem-and-leaf plots of the district-level population means of the covariate X and the outcome Y. The leaves representing the districts in the outlying county are underlined.

X	log(Sk)	Y
977651	9.4	
99998877666533222110000	9.5	
998888887777666554433221000	9.6	47
87543332211000	9.7	001222223456666778889999
	9.8	011122333444555666666777888999999
4	9.9	00122233558
97621	10.0	1
11	10.1	6667
	10.2	066
	10.3	0

To assess the impact of asymmetry, we repeated the simulations with the eight districts of the outlying county removed. The results changed very little. In particular, the overall comparisons of efficiency have the same features as Figure 3 and Table 2: when the population means $\bar{X}_d$ are available (and are used), EB estimators are on average more efficient than their composition counterparts; without $\bar{X}_d$, composition is more efficient on average, and for 80 – 90% of the districts, for settings `A`, `B`, `D` and `a`.

The EB method loses some efficiency over the ideal because the between-district variance σ_B^2 has to be estimated, and the estimation of its MSE is problematic because a random-effects model is used for a problem with fixed effects. Composition loses some efficiency over the ideal because the between-district variance matrix has to be estimated. With EB, appropriateness of the model is an additional concern in practice (though not in this simulation study), and model selection is not compatible with efficiency. We see no way of comparing these drawbacks analytically.

3 Conclusion

Empirical Bayes methods and composition are two general approaches to small-area estimation. Properties of the estimators they yield are difficult to assess analytically because of the complexity of the estimation algorithms and because a set of (D) estimators has to be evaluated. Further, the properties depend on the setting — the covariates used in model fitting, and the (external) population information (district-level means) that are available, as well as on the sampling design.

We have set up and conducted a simulation study for an empirical comparison of the alternative estimators. It was based on an artificial population from which replicate samples are drawn and the sets of alternative estimators evaluated. One set of simulations,

comprising 1000 replications, takes about 12 minutes of CPU time on a laptop computer. Our experience suggests that the number of replicates could be reduced substantially (say 2 – 4 times), because the results for the alternative estimators are highly correlated. The simulations can be extended to settings with $\bar{X}_d$ and $\bar{Z}_d$ estimated from an external source, with their values not established with precision.

The simulations demonstrate that analysts can make choices about small-area estimators that are tailored to the details of the study without having to resort to theoretical derivations that have limited applicability. Of course, the simulations have to be repeated for a range of settings to confirm that the choices are not sensitive to some of the details and to our uncertainty about the features of the population. The simulations can be also used for fine-tuning the software intended for the analysis.

The simulations can be conducted on past data, to establish which method and setting is best suited for the 'current' inference. For example, the methods can be applied to 2001 as the 'past' and 2002 as the 'present', if the appropriate data are available. The best suited method and model is then applied to the current problem (inference about wages in 2006) with the data from 2005 used as the auxiliary information. This approach relies on the realistic assumption that the relevant features and distributions in the population of the country have not been altered substantially in the course of a few years.

All the software was written in `R` (`R` Development Core Team, 2004), and is available from the author on request.

Appendix

We describe here the algorithm for fitting the model in (1) and how the EB estimates for the districts are derived from the model fit. Details of the composite estimator are also given.

The model in (1) is fitted by the Fisher scoring algorithm. It comprises iterations of the following sets of equations:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{new}} &= \left(\mathbf{X}^\top \hat{\mathbf{W}}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^\top \hat{\mathbf{W}}^{-1} \mathbf{y} \\
\hat{\sigma}^2_{\text{new}} &= \frac{1}{n}\, \mathbf{e}^\top \hat{\mathbf{W}}^{-1} \mathbf{e} \\
\hat{\omega}_{\text{new}} &= \hat{\omega}_{\text{old}} + \left\{ \sum_{d=1}^{D} \left(\hat{g}_d\, n_d\right)^2 \right\}^{-1} \left\{ -\sum_{d=1}^{D} \hat{g}_d\, n_d + \frac{1}{\hat{\sigma}^2_{\text{old}}} \sum_{d=1}^{D} \left(\hat{g}_d\, \mathbf{1}^\top \mathbf{e}_d\right)^2 \right\},
\end{aligned}$$

where $\mathbf{X}$ is the regression matrix formed by stacking the rows $\mathbf{x}_{id}$, $\mathbf{W} = \text{diag}_d\,(\mathbf{W}_d)$ is the block-diagonal scaled variance matrix of the outcomes $\mathbf{y}$, composed of the blocks $\mathbf{W}_d = \sigma^{-2}\text{var}(\mathbf{y}_d) = \mathbf{I} + \omega\mathbf{J}$, $\omega = \sigma^2_{\text{B}}/\sigma^2_{\text{W}}$ is the variance ratio, $\mathbf{1}$ is the column vector of ones of appropriate length, $\mathbf{e}_d = \mathbf{y}_d - \mathbf{X}_d\hat{\boldsymbol{\beta}}_{\text{old}}$ and $\hat{g}_d = 1/(1 + n_d\,\hat{\omega}_{\text{old}})$; the subscripts 'old'

and 'new' indicate the estimate after the previous and the current iteration, respectively, and hats indicate estimates.

The iterations converge very quickly; after ten iterations, the changes in the values of all estimated parameters are smaller than 10^{-6}, but often six to eight iterations suffice for both the valid model, which was used for generating the data, and its submodels (see Table 2). The iterations commence with the ordinary least squares solution and the value $\hat{\omega}_0 = 0.25$.

Upon convergence, $\bar{Y}_d$ is estimated by $\hat{\mathbf{X}}_d\hat{\boldsymbol{\beta}} + \hat{\delta}_d$, where $\hat{\delta}_d = \hat{g}_d\,\mathbf{e}_d^\top\mathbf{1}\hat{\omega}$. The vector $\hat{\mathbf{X}}_d$ is composed of the sample means of the regression variables. However, if the district-level population mean of a covariate in $\mathbf{X}$ is known, it is substituted for the estimate. For example, when $\bar{Z}_d$ is known but $\bar{X}_d$ is not, we estimate $\bar{Y}_d$ by $\hat{\beta}_0 + \bar{x}_d\,\hat{\beta}_\mathrm{x} + \bar{Z}_d\,\hat{\beta}_\mathrm{z} + \hat{\delta}_d$ when the regression in (1) is on X and Z, by $\hat{\beta}_0 + \bar{Z}_d\,\hat{\beta}_\mathrm{z} + \hat{\delta}_d$ when the regression is only on Z, and by $\hat{\beta}_0 + \hat{\delta}_d$ when the regression is empty and contains only the intercept. Note that the estimates $\hat{\beta}_0$, $\hat{\beta}_\mathrm{z}$ and $\hat{\delta}_d$ in these expressions differ because they depend on the model that was fitted.

In composite estimation for a district d, we form a vector of district-level quantities $\mathbf{u}_d$ from the sample means $\bar{y}_d$, $\bar{x}_d$ and $\bar{z}_d$ or, in general, estimators of the district-level population means $\bar{Y}_d$, $\bar{X}_d$ and $\bar{Z}_d$, as well as these population means (or percentages) when they are available. Let $\mathbf{u}$ be the national counterpart of the $\mathbf{u}_d$, formed by the estimates or population quantities of the variables in $\mathbf{u}_d$. Let $\mathbf{c}$ be the vector that indicates the estimator $\bar{y}_d$ in $\mathbf{u}_d$, that is, $\bar{y}_d = \mathbf{c}^\top\mathbf{u}_d$. For example, when $\bar{y}_d$ is the first component of $\mathbf{u}_d$, $\mathbf{c} = (1, \ldots, 0)^\top$. We seek the vector of coefficients $\mathbf{b}_d$ such that the multivariate convex combination

$$\tilde{Y}_d \;=\; (\mathbf{c} - \mathbf{b}_d)^\top\mathbf{u}_d + \mathbf{b}_d^\top\mathbf{u}$$

has the smallest MSE. With simple random sampling within districts, the solution would be

$$\mathbf{b}_d^* \;=\; \left(1 - \frac{n_d}{n}\right)\left\{\mathbf{V}_d\left(1 - 2\,\frac{n_d}{n}\right) + \mathbf{V} + \boldsymbol{\Sigma}_\mathrm{B}\right\}^{-1}\mathbf{V}_0\mathbf{c}\,, \tag{2}$$

if the sampling variance matrices $\mathbf{V}_d = \mathrm{var}(\mathbf{u}_d \,|\, d)$ and $\mathrm{var}(\mathbf{u}) = \mathbf{V}$ and the district-level variance matrix $\boldsymbol{\Sigma}_\mathrm{B}$ of the within-district expectations of $\mathbf{u}_d$ were known; see Longford (2005, Chapter 8). Estimation of the sampling variance matrices is a standard task in sampling theory; for population quantities in $\mathbf{u}_d$, the corresponding variances in $\mathbf{V}_d$ and $\mathbf{V}$ vanish.

The matrix $\boldsymbol{\Sigma}_\mathrm{B}$ is estimated by moment matching. First, the within-district variance matrix $\boldsymbol{\Sigma}_\mathrm{W}$ is estimated by pooling the within-district sums of squares and crossproducts. District-level population means do not contribute to $\boldsymbol{\Sigma}_\mathrm{W}$; they are regarded as constants within districts. Then the statistic

$$\mathbf{S} \;=\; \sum_{d=1}^{D} n_d\,(\mathbf{u}_d - \mathbf{u})(\mathbf{u}_d - \mathbf{u})^\top$$

is evaluated and used in the estimator

$$\hat{\mathbf{\Sigma}}_{\mathrm{B}} \;=\; \frac{1}{n - \frac{1}{n}\sum_{d=1}^{D} n_d^2} \left\{ \mathbf{S} - (n-1)\hat{\mathbf{\Sigma}}_{\mathrm{W}} \right\} .$$

The vector of coefficients $\mathbf{b}_d$ is estimated naively, using estimators of $\mathbf{V}_d$, $\mathbf{V}$ and $\mathbf{\Sigma}_{\mathrm{B}}$.

With a general sampling design, the sample quantities in $\mathbf{u}_d$ and $\mathbf{u}$ can be estimated by statistics other than sample means (e.g., by the Horvitz-Thompson estimator), and their sampling variance matrices estimated accordingly. The estimator $\hat{\mathbf{\Sigma}}_{\mathrm{B}}$ also has to be adjusted, but it can always be motivated as an estimator of the excess variation of the estimates $\mathbf{u}_d$, beyond what is expected based on the sampling variance matrices $\mathbf{V}_d$ or their estimates $\hat{\mathbf{V}}_d$. See Longford (2005, Chapter 8) for details.

Acknowledgements

This manuscript was prepared while the author was an academic visitor at the Department of Economics and Business, University of Pompeu Fabra, Barcelona, Spain. Financial support of the Spanish Ministry of Science and Technology, through the Grants No. SEC2003-04476 and SAB2004-0190, is acknowledged.

References

Longford, N. T. (2005). *Missing Data and Small-Area Estimation. Modern Analytical Equipment of the Survey Statistician.* Springer-Verlag, New York.

Longford, N. T. (2006a). Using small-area estimates. *Statistics in Transition* **7**, 715–735.

Longford, N. T. (2006b). On standard errors of small-area estimators. Submitted.

R Development Core Team (2004). R: *A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria,