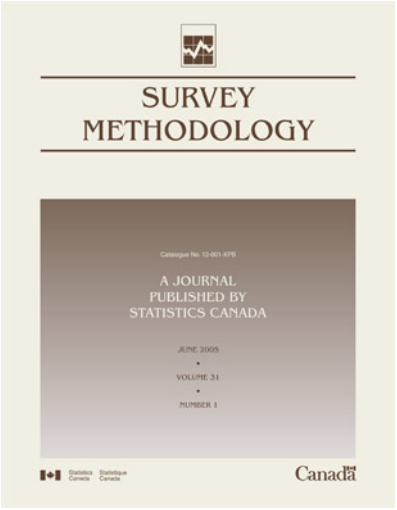
Catalogue no. 12-001-XIE

Survey Methodology

2005



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

2005

Published by authority of the Minister responsible for Statistics Canada

November 2005

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

4
Vol. 29, No. 1, pp. 33-44
Statistics Canada, Catalogue No. 12-001

The Effect of Model Choice in Estimation for Domains, Including Small Domains

Risto Lehtonen, Carl-Erik Särndal and Ari Veijanen [1]

Abstract

In this paper we examine the effect of model choice on different types of estimators for totals of domains, including small domains (small areas), for a sampled finite population. The paper asks: How do different estimator types compare for a common underlying model statement? We argue that estimator type (Synthetic, GREG, Composite, EBLUP, hierarchical Bayes, and so on) is one important aspect of domain estimation, and that the choice of the model, including its parameters and effects, is a second aspect, conceptually different from the first. Earlier work has not always kept this distinction clear. For a given estimator type, one can derive different estimators, depending on the choice of model. A number of estimator types have been proposed in the recent literature, but there is relatively little of an impartial comparison between them. In this paper we discuss three types: Synthetic, GREG, and, to a limited extent, Composite. We show that model improvement (the transition from a weaker to a stronger model) has very different effects on the different estimator types. We also show that the difference in accuracy between the different estimator types depends on the choice of model. For a well-specified model the difference in accuracy between Synthetic and GREG is negligible, but it can be substantial if the model is misspecified. Synthetic then tends to be highly inaccurate. We rely partly on theoretical results (for simple random sampling only), partly on empirical results. The empirical results are based on simulations with repeated samples drawn from two finite populations, one artificially constructed, the other constructed from real data from the Finnish Labour Force Survey.

Key Words: Survey sampling; Generalized regression estimator; Synthetic estimator; Composite estimator; Multi-level models; Small areas; Small domains.

1. Background

Most surveys require that estimates be made not only for the entire population under study but also for a number of sub-populations, called *domains* or *domains of interest*. Estevao and Särndal (1999) give a general outline of estimation for domains from a design-based perspective, with the use of auxiliary information. The sampling design is general, and so is the vector of auxiliary variables. The framework is also called model-assisted. Several national statistical agencies have in recent years constructed software that routinely handles domain estimation within the design-based, model-assisted framework. Examples of such software include CLAN97 by Statistics Sweden and GES by Statistics Canada. In a typical survey, some domains of interest are large enough, and the auxiliary information strong enough, so that the design-based estimators will be sufficiently accurate. But other domains may be so small (contain so few sampled units) that the design-based estimates will be too erratic. The statistical agency may then decide to suppress the publication of statistics for such domains.

Model-dependent estimates are less volatile, but an unattractive feature is their unknown bias, which can be substantial. The model-dependent synthetic estimator has occupied a prominent place in research on small area estimation from around 1970 and on, see for example, National Center for Health Statistics (1968), National Research Council (1980). Different estimators built on nested error regression models (Fuller and Battese 1973), random regression coefficients models (Dempster, Rubin and Tsutakawa 1981) and simple random effects models (Fay and Herriot 1979) provide examples of early propositions for alternatives to the synthetic estimator. Various composite estimators, constructed as weighted combinations of a model dependent estimator and a design-based estimator, were also proposed in the literature (for example Holt, Smith and Tomberlin 1979).

It was in connection with the synthetic estimator that the term "borrowing strength" began to be widely used. Today this term is invoked in virtually every one of the many published articles on small area estimation. Together, these articles now provide a rich source of possibilities for small area estimation, a majority of them model dependent. They draw on a variety of established statistical arguments and principles, such as generalized linear mixed models, composite estimation, empirical Bayes estimation, hierarchical Bayes, and so on.

Borrowing strength (or information) via modeling is a recurring theme in recent literature on small area estimation (for example Ghosh and Rao 1994; Pfeffermann 1999; Rao 1999). Borrowing strength is generally understood to mean that the estimator in use depends on data on the variable of interest, denoted y, from "related areas" or more generally from a larger area, in an effort to improve the accuracy for the small area. The resulting estimator is called *indirect*, in contrast to the one that uses y−data strictly from the domain itself, in which case it is called *direct*.

1. Risto Lehtonen, University of Jyväskylä, Department of Mathematics and Statistics, P.O. Box 35 (MaD), FIN-40014 U. Jyväskylä, Finland; Carl-Erik Särndal, 2115 Ernbrook #44, Ottawa, Ontario, K1B 4J5; Ari Veijanen, Statistics Finland, P.O. Box 4 V, FIN-00022 Statistics Finland, Finland.

Underlying models and their features is another prominent theme in recent literature (for example Ghosh, Natarajan, Stroud and Carlin 1998; Marker 1999; Moura and Holt 1999; Prasad and Rao 1999; Feder, Nathan and Pfeffermann 2000). Small area estimates, and domain estimates more generally, are intrinsically linked to the idea of modeling. Holt and Rao (1995) hint that the use of $y-$information from other areas, although in a sense "necessary", should not be carried to an extreme. Instead there should be "specific allowance for local variation" through a model formulation that includes area-specific effects. This raises a certain ambiguity: borrowing strength from other areas is desirable, even necessary, but only within limits. It is unclear what these limits should be.

There is an extensive recent literature on small area estimation from a Bayesian point of view, including empirical Bayes and hierarchical Bayes techniques (for example Datta, Lahiri, Maiti and Lu 1999; Ghosh and Natarajan 1999; You and Rao 2000). Some recent publications relate frequentist and Bayesian approaches in small area estimation (for example Singh, Stukel and Pfeffermann 1998). Rao (2003) provides a good overview of current literature on model-based small area estimation.

The discussion in recent literature of domain estimation, including small area estimation, revolves around three crucial concepts: (i) borrowing strength; (ii) the type of (implicit or explicit) model, (iii) the parameters or effects admitted in the model statement, that is, whether they should be area specific or defined at some higher level of aggregation such as a set of "similar areas". We agree that these concepts are central and we use them in this paper.

Our starting point for the paper is summarized by (i) to (iii) as follows: (i) a number of different estimator types have been proposed for domain estimation and small area estimation: Synthetic estimator, Generalized Regression (GREG) estimator, Composite estimator, Empirical Best Linear Unbiased Predictor (EBLUP), empirical Bayes (EB) estimator, hierarchical Bayes estimator and so forth; (ii) for every estimator type, different estimators result from the choice of model; (iii) to borrow or not to borrow strength becomes an issue for some of the model choices. Attempts at borrowing strength takes place when the estimation of the parameters and effects in the model requires the use of $y-$values for units outside the domain itself.

2. Statement of Objectives

An objective in this paper is to examine domain estimation through a separation of two ideas: estimator type on the one hand, the choice of the underlying model on the other. We get a two-dimensional arrangement of possible estimators: By estimator type, by model choice. This distinction has not been emphasized enough in earlier literature.

We study the effect of model choice, and of model improvement, on selected estimator types: the Generalized Regression (GREG) estimator (which is design-based), the Synthetic (SYN) estimator (which is model dependent) and the Composite estimator with Empirical Best Linear Unbiased Predictor EBLUP as a special case (which also is model dependent). By construction, each type has its own particular features. For example the GREG estimator type is constructed to be design unbiased, the model dependent ones usually are not. The GREG estimator's variance, although of order n^{-1}, can be very large for a small domain if the "effective sample size" is small; GREG is a "strongly design consistent" estimator in that its relative bias (bias divided by standard deviation) tends to zero as $n^{-1/2}$. The SYN estimator is usually design biased; its bias does not approach zero with increasing sample size; its variance is usually smaller than that of GREG. The EBLUP is design consistent (although not strongly design consistent in the manner of GREG); is design biased for any fixed finite sample size; its variance ordinarily falls between that of GREG and that of SYN.

The chosen model specifies a hypothetical relationship between the variable of interest, y, and the vector of predictor variables, $\mathbf{x}$, and makes assumptions about its perhaps complex error structure. For every specified model, we can derive one GREG estimator, one SYN estimator, one composite estimator, by observing the respective construction principles. An "improved model" will influence all of GREG, SYN and composite, usually so that the MSE decreases. In other words, if Model A is better than Model B, the SYN estimator for Model A is usually better than the SYN estimator for Model B. The same is usually the case for GREG.

Model choice has two aspects: (i) the mathematical form, or the type, of the model, and (ii) the specification of the parameters and effects in the model. For a given variable of interest, some models are more appropriate than others. Model improvement can result either from a more appropriate model type, or from a better parametization, or both. We can distinguish linear models and nonlinear models. Logistic models are a special case of the latter. For a binary or polytomous variable of interest y, a (multinomial) logistic model type is arguably an improvement on a linear model type, because the fitted values under the former will necessary fall in the unit interval, which is not always true for a linear model. Lehtonen and Veijanen (1998) introduced the logistic GREG estimator and studied it in the context of the Finnish Labour Force survey. Another example is when a Bayesian model formulation is preferred to other forms.

The second aspect of model choice is the specification of the parameters and effects in the model. Some of these may be defined at the fully aggregated population level, others at the level of the domain (area specific parameters), yet others at some intermediate level (for a set of "related areas"). Using a multi-level model type, we can introduce stochastic effects that recognize domain differences, as in Goldstein (1995) for the SYN estimator and by Lehtonen and

Veijanen (1999) for the GREG estimator. They found improved accuracy in small domains, compared to the GREG estimator based on a model with fixed effects at the population level. Generally, model improvement occurs when more parameters or effects are added to the model, as for example when it is formulated to include area specific effects reflecting local variation.

We show in this paper (i) that model improvement will generally, for any estimator type considered here, be accompanied by a decrease in MSE; (ii) that the effect on the MSE of model improvement is very different for different estimator types; (iii) that for a well-specified model, there are negligible differences only in the accuracy (the MSE) achieved by the estimator types under study, but under model failure the differences can be substantial. We emphasize that a comparison of estimators of different types should only take place under "similar conditions". That is, the model choice must be the same for all alternatives considered. An estimator is shown to be better than another estimator only if the MSE of the former is smaller than that of the latter, for one and the same model choice. (It is difficult to establish that one estimator type is uniformly better than another, that is, better under all model choices.)

Table 1 shows the estimators to be discussed, in a two-way arrangement by estimator type and by model choice. This table also shows our notation for the estimators to be considered. There are six SYN type estimators and six GREG type estimators in the table. Each of the six rows corresponds to a different model choice. A population model (P−model; rows 1 and 2) is one whose only parameters are fixed effects defined at the population level; it contains no domain specific parameters. A domain model (D−model) is one having at least some of its parameters or effects defined at the domain level. These are fixed effects for rows 3 and 4, or mixed with random effects for rows 5 and 6. "Linear" and "logistic" refer to the mathematical form. In this paper we discuss all estimators in Table 1 except the two in the last row.

Table 1
Schematic Presentation of the SYN and GREG Estimators by Model Choice and Estimator Type

Model choice			Estimator type: Model-dependent synthetic	Estimator type: Model-assisted generalized regression
Fixed-effects models	Population models	Linear	SYN−P	GREG−P
		Logistic	LSYN−P	LGREG−P
	Domain models	Linear	SYN−D	GREG−D
		Logistic	LSYN−D	LGREG−D
Mixed models including fixed and random effects	Domain models	Linear	MSYN−D	MGREG−D
		Logistic	MLSYN−D	MLGREG−D

In addition to the SYN and GREG estimator types listed in Table 1, we can consider composite estimators of the type $\hat{\gamma}_d \mathrm{GREG} + (1-\hat{\gamma}_d)\mathrm{SYN}$, being appropriately weighted combinations of the corresponding GREG and SYN estimators. In this paper we examine one estimator of this type, the EBLUP estimator.

The paper is organized as follows: Section 3 introduces three types of estimators for a domain total. In section 4, we describe the models used in the construction of these estimators. In section 5 we derive analytically the effect of model improvement, in a simple case. (Only simple cases can be treated analytically, because the formulas quickly attain a high degree of complexity, depending on the sampling design and other factors.) Section 6 is devoted to Monte Carlo simulations for two finite populations, illustrating the effect of model improvement on the three selected estimator types. Summary and discussion is given in section 7.

3. Estimators of Domain Totals

The finite population is denoted $U = \{1, 2, ..., k, ..., N\}$. A probability sample *s* is drawn from *U* by a given sampling design such that unit *k* is given the inclusion probability π_k. The sampling weight of unit *k* is then $a_k = 1/\pi_k$. Denote by *y* the variable of interest and by y_k its value for unit *k*. We consider a set of mutually exhaustive and exhaustive domains $U_1, ..., U_d, ..., U_D$. The target parameters are the set of domain totals, $Y_d = \sum_{U_d} y_k$, $d = 1, ..., D$.

Auxiliary information is essential for building accurate domain estimators, and increasingly so when domains of interest get smaller. Let $\mathbf{x}$ be the auxiliary vector of dimension $J \geq 1$ with a known value $\mathbf{x}_k$ for every unit $k \in U$. In a survey on individuals, $\mathbf{x}_k$ may specify known data about person *k*, such as age class, sex and other continuous or qualitative variable values. We assume that the vector value $\mathbf{x}_k$ and domain membership is known and specified in the frame for every $k \in U$. (For some estimators, it suffices to know the *total* of $\mathbf{x}_k$ for each domain of interest.)

The estimators we consider are constructed as follows: The first step is to estimate the designated model, using the sample data $\{(y_k, \mathbf{x}_k); k \in s\}$. Next, using the estimated parameter values, the vector value $\mathbf{x}_k$ and the domain membership of *k*, we compute the predicted value $\hat{y}_k$ for every $k \in U$, which is possible under our assumptions because $\mathbf{x}_k$ is known for every $k \in U$. The predictions, $\{\hat{y}_k; k \in U\}$, and the observations, $\{y_k; k \in s\}$, provide the material for the estimator types considered here.

Consider a fixed-effects model specification, linear or nonlinear, such that $\mathrm{E}_m(y_k) = f(\mathbf{x}_k; \beta)$, for a given function $f(\cdot; \beta)$, where β is an unknown parameter vector requiring estimation, and E_m refers to the expectation under the model. The model fit yields the estimate $\hat{\beta}$. The supply of predicted values $\hat{y}_k = f(\mathbf{x}_k; \hat{\beta})$ is computed for $k \in U$. Similarly, for a linear mixed model involving random effects in addition to the fixed effects, the model specification is $\mathrm{E}_m(y_k \mid \mathbf{u}_d) = \mathbf{x}_k'(\beta + \mathbf{u}_d)$ where $\mathbf{u}_d$ is a vector of random effects defined at the domain level. Using

the estimated parameters, predicted values $\hat{y}_k = \mathbf{x}_k'(\hat{\beta} + \hat{\mathbf{u}}_d)$ are computed for all $k \in U$. The models used in this paper are described in more detail in section 4. In more general terms, the models for the construction of GREG and SYN type estimators of domain totals are often members of the family of generalized linear mixed models (for example McCullogh and Searle 2001).

The predictions $\{\hat{y}_k; k \in U\}$ differ from one model specification to another. For a given model specification, the estimator of the domain total $Y_d = \sum_{U_d} y_k$ has the following structure for the three estimator types (Synthetic, Generalized Regression, Composite) to be studied:

$$\hat{Y}_{d\text{SYN}} = \sum_{U_d} \hat{y}_k \tag{3.1}$$

$$\hat{Y}_{d\text{GREG}} = \sum_{U_d} \hat{y}_k + \sum_{s_d} a_k (y_k - \hat{y}_k) \tag{3.2}$$

$$\hat{Y}_{d\text{COMP}} = \sum_{U_d} \hat{y}_k + \hat{\gamma}_d \sum_{s_d} a_k (y_k - \hat{y}_k) \tag{3.3}$$

where $a_k = 1/\pi_k$, $s_d = s \cap U_d$ is the part of the full sample s that falls in U_d, and $d = 1, ..., D$. $\hat{Y}_{d\text{SYN}}$ relies heavily on the truth of the model, and is usually biased. On the other hand, $\hat{Y}_{d\text{GREG}}$ has a second term that protects against model misspecification. The domain-specific weight $\hat{\gamma}_d$ in $\hat{Y}_{d\text{COMP}}$ is appropriately constructed to meet certain optimality properties, as explained in section 6. The weight $\hat{\gamma}_d$ approaches unity for increasingly large domain sample sizes, so that $\hat{Y}_{d\text{COMP}}$ approaches $\hat{Y}_{d\text{GREG}}$. At the other extreme, when $\hat{\gamma}_d$ is near zero, $\hat{Y}_{d\text{COMP}}$ is close to $\hat{Y}_{d\text{SYN}}$. We note that for a given model specification, (3.2) and (3.3) reduce to (3.1) for a domain d with no sample elements in s_d.

4. Models

4.1 Fixed-Effects Linear Models

Let $\mathbf{x}_k = (1, x_{1k}, ..., x_{jk}, ..., x_{Jk})'$ be a $(J+1)-$dimensional vector containing the values of $J \geq 1$ predictor variables x_j, $j = 1, ..., J$. This vector is used to create the predicted values $\hat{y}_k$ in the estimators (3.1), (3.2) and (3.3).

The estimators SYN−P and GREG−P build on the model specification (called the P−model)

$$\mathrm{E}_m(y_k) = \mathbf{x}_k'\beta \tag{4.1}$$

for $k \in U$, where $\beta = (\beta_0, \beta_1, ..., \beta_J)'$ is a vector of fixed effects defined for the whole population. If $y-$data were observed for the whole population, we could compute the generalized least squares (GLS) estimator of β given by

$$\mathbf{B} = \left(\sum_U \mathbf{x}_k \mathbf{x}_k' / c_k\right)^{-1} \sum_U \mathbf{x}_k y_k / c_k \tag{4.2}$$

where the c_k are specified positive weights. With no significant loss of generality we specify these to be of the form $c_k = \lambda' \mathbf{x}_k$ for $k \in U$, where the $(J+1)-$vector λ does not depend on k. Because (4.2) cannot be computed, the fit is carried out in practice on the observed sample data, yielding

$$\hat{\mathbf{B}} = \left(\sum_s a_k \mathbf{x}_k \mathbf{x}_k' / c_k\right)^{-1} \sum_s a_k \mathbf{x}_k y_k / c_k \tag{4.3}$$

where $a_k = 1/\pi_k$ is the sampling weight of unit k. The resulting predicted values are $\hat{y}_k = \mathbf{x}_k' \hat{\mathbf{B}}$. They can be computed for all $k \in U$.

The estimators SYN−D and GREG−D are built with the same predictor vector $\mathbf{x}_k$, but with an improved model specification (called the D−model) allowing a fixed-effects vector β_d separately for every domain, so that

$$\mathrm{E}_m(y_k) = \mathbf{x}_k' \beta_d \tag{4.4}$$

for $k \in U_d$, $d = 1, ..., D$, or equivalently,

$$\mathrm{E}_m(y_k) = \sum_{d=1}^{D} \delta_{dk} \mathbf{x}_k' \beta_d \tag{4.5}$$

for $k \in U$, where δ_{dk} is the domain indicator of unit k, defined by $\delta_{dk} = 1$ for all $k \in U_d$, and $\delta_{dk} = 0$ for all $k \notin U_d$, $d = 1, ..., D$. If the model (4.3) could be fitted to data for the whole population, the GLS estimator of β_d would be

$$\mathbf{B}_d = \left(\sum_{U_d} \mathbf{x}_k \mathbf{x}_k' / c_k\right)^{-1} \sum_{U_d} \mathbf{x}_k y_k / c_k. \tag{4.6}$$

In practice, the fit must be based on the observed sample data, leading to

$$\hat{\mathbf{B}}_d = \left(\sum_{s_d} a_k \mathbf{x}_k \mathbf{x}_k' / c_k\right)^{-1} \sum_{s_d} a_k \mathbf{x}_k y_k / c_k. \tag{4.7}$$

The resulting predicted values are given by $\hat{y}_k = \mathbf{x}_k' \hat{\mathbf{B}}_d$ for $k \in U_d$, $d = 1, ..., D$. Because of the specification $c_k = \lambda' \mathbf{x}_k$, we have $\sum_{s_d} a_k (y_k - \hat{y}_k) = 0$. Consequently, SYN−D and GREG−D are identical, that is, $\hat{Y}_{d\text{SYN}-D} = \hat{Y}_{d\text{GREG}-D}$ for every sample s.

The transition from GREG−P to GREG−D, and from SYN−P to SYN−D, affects the MSE in a way to be analyzed in section 5. SYN−P and GREG−P will be examined empirically in section 6.

4.2 Linear Mixed Models

The estimators MSYN−D and MGREG−D build on a two-level linear model (called the D−model) involving fixed as well as random effects recognizing domain differences,

$$\begin{aligned} \mathrm{E}_m(y_k \mid \mathbf{u}_d) &= \beta_0 \\ &\quad + u_{0d} + (\beta_1 + u_{1d}) x_{1k} \\ &\quad + ... + (\beta_J + u_{Jd}) x_{Jk} \\ &= \mathbf{x}_k'(\beta + \mathbf{u}_d) \end{aligned} \tag{4.8}$$

for $k \in U_d$, $d = 1, ..., D$. Each coefficient is the sum of a fixed component and a domain specific random component: $\beta_0 + u_{0d}$ for the intercept and $\beta_j + u_{jd}$, $j = 1, ..., J$ for the slopes. The components of $\mathbf{u}_d = (u_{0d}, u_{1d}, ..., u_{Jd})'$

represent deviations from the coefficients of the fixed-effects part of the model,

$$\mathrm{E}_m(y_k) = \beta_0 + \beta_1 x_{1k} + ... + \beta_J x_{Jk} = \mathbf{x}_k'\beta, \qquad (4.9)$$

which agrees with (4.1). More generally, we can have that only some of the coefficients in (4.8) are treated as random, so that, for some j, $u_{jd} = 0$ for every d. One of the simplest special cases of (4.8), commonly used in practice, is the one that includes a domain-specific random intercept u_{0d} as the only random term, as in one of the models used in section 6. Another model used in section 6 is the special case of (4.8) for $J = 1$, with a random slope u_{1d} and a random intercept u_{0d}.

We insert the resulting fitted $y-$values, $\hat{y}_k = \mathbf{x}_k'(\hat{\beta} + \hat{\mathbf{u}}_d)$, into (3.1) to obtain the two-level MSYN−D estimator. Inserting the fitted values, $\hat{y}_k = \mathbf{x}_k'(\hat{\beta} + \hat{\mathbf{u}}_d)$, into (3.2), we obtain the two-level MGREG−D estimator, introduced by Lehtonen and Veijanen (1999). MSYN−D and MGREG−D will be examined empirically in section 6.

For the simulations reported in section 6, we fitted the two-level model (4.8) by the iterative least squares fitting (IGLS) algorithm of Goldstein (1995). Random effects were estimated by equation (2.2.2) in Goldstein (1995). This algorithm appeals to an assumption that the random effects follow a joint normal distribution $N(\mathbf{0}, \Omega)$. Note however that this assumption of normality is in no way necessary to obtain favorable properties for the resulting MGREG−D estimator. It is nearly unbiased regardless of any such assumption. The fitting of a multi-level model is more demanding than the fitting of a linear fixed-effects model, since estimation of the covariance matrix Ω is required.

4.3 Logistic Models

The estimators LSYN−P and LGREG−P build on a multinomial logistic P−model. Assume an $m-$class polytomous response defined by the class variables y_i with value $y_{ik} = 1$ if k belongs to class i and $y_{ik} = 0$ otherwise, $i = 1, ..., m,$ and modeled by

$$E_m(y_{ik}) = \frac{\exp(\mathbf{x}_k'\beta_i)}{\sum_{r=1}^{m}\exp(\mathbf{x}_k'\beta_r)} \qquad (4.10)$$

for $k \in U$, where $\mathbf{x}_k = (1, x_{1k}, ..., x_{jk}, ..., x_{Jk})'$ and $\beta_i = (\beta_{i0}, \beta_{i1}, ..., \beta_{iJ})'$ are vectors of fixed effects defined for whole population. To avoid identifiability problems, we set $\beta_1 = 0$. The LSYN−P and LGREG−P estimators of the population frequency of class i in domain d, $Y_{id} = \sum_{U_d} y_{ik}$, are defined by (3.1) and (3.2), respectively, if we replace y_k and $\hat{y}_k$ by y_{ik} and $\hat{y}_{ik} = \exp(\mathbf{x}_k'\hat{\beta}_i) / (1 + \sum_{r=2}^{m}\exp(\mathbf{x}_k'\hat{\beta}_r))$, where $\hat{\beta}_i$ is the estimate of β_i obtained from the fit of (4.10).

LGREG−P was introduced and studied in Lehtonen and Veijanen (1998). LSYN−P and LGREG−P will be examined empirically in section 6, where $\hat{\beta}_i$ is derived as a pseudo-maximum likelihood estimator incorporating the sampling weights.

5. Analytic Examination of the Effect of Model Improvement

In this section we analyze the transition from GREG−P to GREG−D, and from SYN−P to SYN−D in the case of Simple Random Sampling. For both estimator types, GREG and SYN, we find that the accuracy is improved when the model changes from the weaker P−model (4.1) (with fixed effects at the level of the whole population) to the stronger D−model (4.5) (admitting fixed effects at the domain level). Intuitively, this is to be expected. What is of interest here is the pattern of improvement. It is very different for the two types.

Our objective is to measure the effect of model improvement on $\hat{Y}_d$, where $\hat{Y}_d$ denotes either $\hat{Y}_{d\mathrm{GREG}}$ or $\hat{Y}_{d\mathrm{SYN}}$. For this purpose, we use the relative improvement in MSE,

$$\mathrm{RELIMP}(\hat{Y}_d) = (\mathrm{MSE}_{dP} - \mathrm{MSE}_{dD}) / \mathrm{MSE}_{dD} \qquad (5.1)$$

where MSE_{dP} and MSE_{dD} denote the MSE of $\hat{Y}_d$ under the P−model and under the D−model, respectively. Both MSE_{dP} and MSE_{dD} depend on the sampling design and on the composition of the $\mathbf{x}_k-$vector. The improvement factor (5.1) is in general a complex formula. It lends itself to easy analytic interpretation only in simple cases. Therefore, we examine here the case of Simple Random Sampling Without Replacement (SRS). For other designs and model formulations, empirical studies are necessary. One such study is reported in section 6.

We use the improvement factor (5.1) to measure the effect of changing from the P−model (4.1) (the weaker model) to the D−model (4.5) (the stronger model). The Technical Appendix gives the necessary expressions for bias and MSE of GREG and SYN estimators in the case of an SRS sample of size n from U. The size, n_d, of the sample from the domain U_d is random with expected value $nP_d = nN_d / N$. For GREG, we use (A.5) in Technical Appendix, and the two different forms of E_k presented there, to arrive at

$$\mathrm{RELIMP}(\hat{Y}_{d\mathrm{GREG}}) = \frac{S^2_{E_P U_d}}{S^2_{E_d U_d}} - 1 + (1 - P_d)\frac{\overline{E}^2_{PU_d}}{S^2_{E_d U_d}}$$

$$\approx (1 - P_d)\frac{\overline{E}^2_{PU_d}}{S^2_{E_d U_d}} \qquad (5.2)$$

where $S^2_{E_d U_d} = (1/(N_d - 1))\sum_{U_d} E^2_{dk}$ and $S^2_{E_P U_d} = (1/(N_d - 1))$ $\sum_{U_d}\{E_{Pk} - \overline{E}_{PU_d}\}^2$ with $\overline{E}_{PU_d} = \sum_{U_d} E_{Pk} / N_d$. (Note that

$\bar{E}_{dU_d} = \sum_{U_d} E_{dk} / N_d = 0$). Similarly, for SYN, we use (A.6) in Technical Appendix, and the two different expressions for E_k presented there, to arrive at

$$\mathrm{RELIMP}(\hat{Y}_{d\mathrm{SYN}}) = \frac{S^2_{(R_d E_P)U}}{P_d S^2_{E_d U_d}} - 1 + \frac{nP_d}{1-f}\frac{\bar{E}^2_{PU_d}}{S^2_{E_d U_d}} \approx \frac{nP_d}{1-f}\frac{\bar{E}^2_{PU_d}}{S^2_{E_d U_d}} \qquad (5.3)$$

where $S^2_{(R_d E_P)U} = (1/(N-1))\sum_U (R_{dk} E_{Pk})^2$. The approximation in (5.3) is a result of keeping only the term proportional to the total sample size *n*. By comparison, the other terms are negligible. The approximation in (5.3) is adequate in many cases, although the deleted part is not always insignificant. Comparing the improvement factors (5.2) and (5.3), we note:

(i) **Improvement factor as a function of the bias.** Comparing (5.2) and (5.3), we see that the improvement of SYN is large compared to that of GREG. The main reason is that SYN is handicapped, under the P−model, by an often considerable squared bias term. As the model improves, this handicap is greatly reduced. At the same time the variance term may increase moderately, so that, somewhat paradoxically, SYN becomes more volatile when the model is improved. For GREG, some improvement occurs when the model improves, as a result of a somewhat reduced variance. The improvement is small, compared to the dramatic improvement of SYN.

(ii) **Improvement factor as a function of domain size.** Suppose that $\bar{E}^2_{PU_d} / S^2_{E_d U_d}$ is constant for all domains. Then, the presence of the relative domain size P_d in (5.3) shows that $\hat{Y}_{d\mathrm{SYN}}$ improves more in larger domains than in small domains (where the need for accuracy improvement is relatively greater). For $\hat{Y}_{d\mathrm{GREG}}$, the pattern is more natural in that the improvement is more pronounced for the smaller domains, due to the factor $(1-P_d)$ in (5.2). But if $\bar{E}^2_{PU_d} / S^2_{E_d U_d}$ varies considerably between domains, these conclusions would be modified.

To throw further light on the generally complex improvement factors (5.2) and (5.3), consider the simple specification $\mathbf{x}_k = 1 = c_k$ for all *k*. Then $\hat{Y}_{d\mathrm{SYN}-P} = N_d \bar{y}_s$, $\hat{Y}_{d\mathrm{GREG}-P} = N_d \bar{y}_{s_d} - (1/f)(n_d - nP_d)\bar{y}_s$ with $f = n/N$ and $\hat{Y}_{d\mathrm{SYN}-D} = \hat{Y}_{d\mathrm{GREG}-D} = N_d \bar{y}_{s_d}$. (Overbar denotes the arithmetic mean over the set defined by the subscript.) Using $(N_d - 1)/(N-1) \approx N_d / N$, we get

$$\mathrm{RELIMP}(\hat{Y}_{d\mathrm{GREG}}) \approx (1-P_d)\frac{(\bar{y}_{U_d} - \bar{y}_U)^2}{S^2_{yU_d}} \qquad (5.4)$$

$$\mathrm{RELIMP}(\hat{Y}_{d\mathrm{SYN}}) \approx P_d \frac{S^2_{yU}}{S^2_{yU_d}} - 1 + \frac{nP_d}{1-f}\frac{(\bar{y}_{U_d} - \bar{y}_U)^2}{S^2_{yU_d}} \approx \frac{nP_d}{1-f}\frac{(\bar{y}_{U_d} - \bar{y}_U)^2}{S^2_{yU_d}} \qquad (5.5)$$

where S^2_{yU} and $S^2_{yU_d}$ are the variances of y_k over U and U_d, respectively. The patterns are now very clear. The term $(\bar{y}_{U_d} - \bar{y}_U)^2 / S^2_{yU_d}$ is present in both expressions. For SYN, we see from (5.5) that the improvement factor is proportional to the whole sample size *n*, hence it can be very large. For GREG, the improvement (5.4) is very small by comparison. If $(\bar{y}_{U_d} - \bar{y}_U)^2 / S^2_{yU_d}$ is constant over all domains, GREG is improved more in smaller domains than in larger ones. The opposite holds for SYN.

The results in this section are limited by the complexity of the analytic expressions. Nevertheless they set the pattern for more general situations now to be studied by empirical examination. As the model improves, we can expect SYN to undergo a very large improvement, in terms of reduced MSE, compared to GREG.

6. Empirical Examination of the Effect of Model Improvement by Monte Carlo Experiments

6.1 Experiments and Monte Carlo Summary Measures

The data for Experiment 1, presented in section 6.2, was generated entirely from a specified model, so it has no basis in any real data. For the 100 domains of this data set we compared the SYN estimator type (3.1) and the GREG estimator type (3.2) under different choices of model for a continuous variable of interest. We fitted a fixed-effects linear model (which created SYN−P and GREG−P estimators) and compared the results with those obtained from the fitting of a two-level linear model (which created MSYN−D and MGREG−D estimators).

In constructing the population for Experiment 2, presented in section 6.3, we took real data on ILO unemployment from Finland's Labour Force Survey (LFS) as a starting point for creating a larger artificial population with 84 regional domains. There, the variable of interest is binary (unemployed or not). We fitted, in addition to a fixed-effects linear model (which created SYN−P and GREG−P estimators) and a two-level linear model (which created MSYN−D and MGREG−D estimators), a fixed-effects binomial logistic model (which created LSYN−P and LGREG−P estimators). For this experiment we also constructed a composite estimator (3.3) as a weighted combination of GREG and SYN estimators, creating a COMP−D estimator.

In Experiments 1 and 2, by using estimates $\hat{Y}_d(s_v)$ from repeated samples s_v; $v = 1, 2, ..., K$, we computed for

each domain $d = 1, ..., D$ the following Monte Carlo summary measures of bias, accuracy and relative improvement in MSE. We use two measures of accuracy, the relative root mean squared error (RRMSE) and the median absolute relative error (MdARE). For Experiment 1, where the response variable is continuous, these two measures give the same message about the accuracy. But for Experiment 2, where the response variable is binary, there is sometimes a difference in the conclusions drawn from the two measures.

(i) Absolute relative bias (ARB), defined as the ratio of the absolute value of bias to the true value:

$$\left| \frac{1}{K} \sum_{v=1}^{K} \hat{Y}_d(s_v) - Y_d \right| \Big/ Y_d . \qquad (6.1)$$

(ii) Relative root mean squared error (RRMSE), defined as the ratio of the root MSE to the true value:

$$\sqrt{\frac{1}{K} \sum_{v=1}^{K} (\hat{Y}_d(s_v) - Y_d)^2} \Big/ Y_d . \qquad (6.2)$$

(iii) Median absolute relative error (MdARE), defined as follows. For each simulated sample s_v; $v = 1, 2, ..., K$, the absolute relative error is calculated and a median is taken over the K samples in the simulation:

$$\underset{\text{over } v = 1, ..., K}{\text{Median}} \{ | \hat{Y}_d(s_v) - Y_d | / Y_d \}. \qquad (6.3)$$

(iv) RELIMP, the relative improvement in MSE, defined in the manner of (5.1).

6.2 Experiment 1: Data Generated from a Model

Monte Carlo design

We used the two-level D−model (4.8) with $J = 1$ to generate an artificial population of one million elements distributed on 100 domains. The elements were randomly allocated to a set of 100 domains with probabilities proportional to $\exp(p_d)$ where p_d follows a uniform distribution in (−3.3). In the generation of values for the $x-$variable and $y-$variable in the d^{th} domain, $d = 1, ..., 100$, we operated in the following way. First, the values of the $x-$variable were obtained as independent realizations of $N(\mu_d, \sigma_d^2)$, where the domain-specific parameters (μ_d, σ_d^2) had first been generated from a bi-variate uniform distribution over $(5.15) \times (15.35)$. Then, the response variable values y_k were generated as

$$y_k = \beta_0 + u_{0d} + (\beta_1 + u_{1d}) x_k + \varepsilon_k \qquad (6.4)$$

with $\beta_0 = 10$ and $\beta_1 = 0.6$. In (6.4), the values of ε_k are independent realizations of $N(0, 1)$, and the random effects u_{0d} and u_{1d} were realized from a bivariate normal distribution with $u_{0d} \sim N(0, 4)$, $u_{1d} \sim N(0, 0.01)$, $d = 1, ..., 100$. We report results for two values of the correlation of the random effects: (a) $\text{Corr}(u_{0d}, u_{1d}) = 0$, and (b) $\text{Corr}(u_{0d}, u_{1d}) = -0.5$. One case of a positive correlation, 0.5, was also studied but the results were similar with those in the zero correlation case and are thus omitted.

We examined four estimators: MSYN−D and MGREG−D based on the two-level D−model (4.8), $y_k = \beta_0 + u_{0d} + (\beta_1 + u_{1d}) x_{1k} + \varepsilon_k$, and SYN−P and GREG−P based on the fixed-effects P−model (4.9), that is, $y_k = \beta_0 + \beta_1 x_k + \varepsilon_k$. Both sets of SYN and GREG estimators were calculated in the zero correlation and negative correlation cases. The conditions are thus ideal for MSYN−D and MGREG−D in the sense that the population follows exactly the model that lies behind these two estimators.

From the generated population we drew $K = 1{,}000$ samples, each of size $n = 10{,}000$, with Simple Random Sampling Without Replacement (SRS). For each estimator and for each domain, we computed the Monte Carlo summary measures of bias, accuracy and relative improvement in MSE in the manner described in (6.1), (6.2), (6.3) and (5.1). The Monte Carlo measures were then averaged with respect to a classification of the domains into Small (25 domains with average domain sample size < 10), Medium-sized (50 domains with average domain sample size ≥ 10 and < 50), and Large (25 domains with average domain sample size ≥ 50).

Results

The results for the cases of zero correlation (a) and negative correlation (b) are given in Tables 2 and 3. In both cases, SYN−P has a large bias (measured by the average ARB) for all the three domain size categories (Table 2). The bias is slightly larger in the zero correlation case. The bias in SYN−P is considerably reduced by MSYN−D, but is still significant in small domains. In the smallest domains, the estimated residuals (the estimates of the random effects) were biased towards zero, which created some bias in the estimates. The accuracy (measured by the average RRMSE and the average MdARE) of MSYN−D (based on the "ideal model") is much better than that of SYN−P (which is based on a population model). Accuracy gains are larger for the zero correlation case, and gains are substantial especially in larger domains. This result is in line with our theoretical results in section 5.

GREG−P and MGREG−D are essentially unbiased, confirming theory. Out of these two, accuracy is clearly better for MGREG−D, especially in small domains. In larger domains, accuracy gains are much smaller for the GREG estimator type than for the SYN estimator type. Bias and accuracy of GREG estimators are quite similar in both zero correlation and negative correlation cases.

Table 2
Average Absolute Relative Biais (ARB) (%), Average Relative Root Mean Squared Error (RRMSE) (%) and Average Median Absolute Relative Error (MdARE) (%) of Total Estimators in Small, Medium-Sized and Large Domains of a Synthetic Population with (a) Random Slope and Intercept Independent or (b) Random Slope and Intercept Negatively Correlated

	Average ARB (%) Expected domain size in sample			Average RRMSE (%) Expected domain size in sample			Average MdARE (%) Expected domain size in sample		
	Small (1 – 9)	Medium (10 – 49)	Large (50+)	Small (1 – 9)	Medium (10 – 49)	Large (50+)	Small (1 – 9)	Medium (10 – 49)	Large (50+)
(a) Zero correlation									
Model-dependent SYN estimators									
SYN–P	10.29	12.37	10.54	10.3	12.4	10.6	10.3	12.4	10.5
MSYN–D	1.32	0.09	0.01	4.7	1.1	0.4	2.6	0.7	0.2
Model-assisted GREG estimators									
GREG–P	0.21	0.06	0.01	7.5	2.5	0.8	5.0	1.7	0.5
MGREG–D	0.83	0.03	0.01	4.8	1.1	0.4	2.7	0.7	0.2
(b) Negative correlation (–0.5)									
Model-dependent SYN estimators									
SYN–P	7.92	9.51	8.26	7.9	9.5	8.3	7.9	9.5	8.3
MSYN–D	1.20	0.09	0.01	4.2	1.1	0.4	2.5	0.7	0.2
Model-assisted GREG estimators									
GREG–P	0.18	0.05	0.01	6.4	2.1	0.6	4.2	1.4	0.4
MGREG–D	0.67	0.02	0.01	4.4	1.1	0.4	2.6	0.7	0.2

As the theoretical discussion in section 5 has also suggested, the effect on the SYN estimator type of model improvement depends strongly on the size of the domain. This is confirmed here: The D–model leads to a considerable MSE improvement (measured by the average RELIMP) for SYN. The improvement is striking for the large domains (Table 3). By contrast, the effect on the GREG estimator type of model improvement is modest, by comparison, and essentially independent of the domain size, as also suggested by the theoretical results.

Table 3
Average Relative Improvement in MSE (%) of Total Estimators in Small, Medium-Sized and Large Domains of a Synthetic Population with (a) Random Slope and Intercept Independent or (b) Random Slope and Intercept Negatively Correlated

Average relative improvement in MSE (%)	Expected domain size in sample		
	Small (1 – 9)	Medium (10 – 49)	Large (50+)
(a) Zero correlation			
MSYN–D versus SYN–P	8.3	332.5	1,278.3
MGREG–D versus GREG–P	1.9	6.0	3.7
(b) Negative correlation (–0.5)			
MSYN–D versus SYN–P	5.1	197.0	734.7
MGREG–D versus GREG–P	1.3	3.6	2.3

The reason for an improved behavior of SYN and GREG estimators is that a two-level (or more generally, a multilevel) model, because of the presence of domain parameters, produces fitted values $\hat{y}_k$ that are on the average closer to the (unobserved) y_k than those obtained by fitting simply the fixed part of the model. In addition, since MSYN–D takes domain differences into account, it is expected to be less biased than the SYN–P estimator based on the fixed part of the two-level model. Still, we find that the MSYN–D estimator has a significant bias, particularly in the smallest domains, for which the estimated random effects tend to be biased towards zero, which pulls the fitted values in the direction of those of the fixed part of the model. MSYN–D and MGREG–D estimators do not differ considerably in their accuracy, even in small domains.

6.3 Experiment 2: Data Adapted from Finland's Labour Force Survey

Monte Carlo design

The empirical data for our Experiment 2 came from the Finnish Labour Force Survey (LFS), conducted monthly by Statistics Finland. Details on the design and the estimation procedure of the LFS are described in Djerf (1997). In this experiment, we estimate the number of unemployed in 84 administrative regions of Finland, based on the NUTS4 classification (European Union's Nomenclature of Territorial Units for Statistics).

To emulate the sampling design of the Finnish LFS, in a fairly realistic manner, we generated a large artificial population by expanding a one-quarter sample data set of the Finnish LFS. The original data set of 32,564 individuals

contained 29,024 respondents. The respondents were replicated by Simple Random Sampling With Replacement until we had reached a total of 3 million records approximating the size of the labour force in Finland.

The variable of interest, *y*, was a binary variable describing whether a person was unemployed of not. In LFS, the definition of unemployment is based on the ILO (International Labour Organisation) concept. Our population data included four auxiliary variables available from administrative registers (and used by Statistics Finland in their LFS): age, sex, region (NUTS2 level regional unit) and the job-seeker indicator, which is a dichotomous indicator showing whether or not a person is registered as an unemployed job-seeker in the administrative records of Finland's Ministry of Labour. Indicator variables were used for 6 age-by-sex classes (3 age groups, 2 sexes). These register-based data were merged with the survey data at the micro level by using personal identification numbers, which are unique in both data sources.

We examined seven estimators. Three model choices were used. First, we constructed the estimators (3.1) and (3.2), based on the linear fixed-effects P−model (4.9) incorporating the main effects for variables age, sex, region and the job-seeker indicator. The model also incorporates the two-variable interaction of age with the job-seeker indicator. The variables and terms in the model were selected in an exploratory data analysis. The resulting domain total estimators are SYN−P and GREG−P.

Secondly, we constructed the estimators (3.1) and (3.2) based on a binomial logistic model (4.10) involving the same model structure as the P−models for SYN−P and GREG−P. The resulting estimators are LSYN−P and LGREG−P.

Thirdly, we constructed the estimators (3.1) and (3.2) based on the two-level D−model (4.8) again involving the same structure in the fixed part as the previous models. The random component of the model, recognizing domain differences, consisted of random intercepts at the domain (NUTS4) level. The resulting estimators are MSYN−D and MGREG−D. For this model choice, we also constructed the composite estimator (3.3). The resulting estimator is denoted by COMP−D. The weight $\hat{\gamma}_d$ in COMP−D was computed as $\hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2/n_d)$, where $\hat{\sigma}_u^2$ and $\hat{\sigma}_\varepsilon^2$ are sample based estimates for unknown parameters in the model's error structure (Ghosh and Rao 1994). The COMP−D estimator is perhaps best described as a pseudo EBLUP (Prasad and Rao 1999), by the fact that the residuals $y_k - \hat{y}_k$ are sample weighted. (A more conventional EBLUP uses unweighted residuals.)

We carried out four independent Monte Carlo experiments. In each experiment, we drew from the generated LFS population $K = 1{,}000$ samples, each of size $n = 12{,}000$ individuals, with SRS. We generated non-response in each sample using a model for the non-response. We modeled the non-response by a logistic model incorporating the same auxiliary variables as the LGREG−P model. The non-response probabilities were estimated from each sample, and the sampling weights were adjusted accordingly. For each estimator and for each domain, we computed the Monte Carlo summary measures defined in section 6.1. These measures were then averaged with respect to a classification of the domains into Small (32 domains with average domain sample size < 60) and Large (52 domains with average domain sample size ≥ 60). We finally averaged these figures over the four experiments.

Results

Table 4 shows the results for the seven estimators. In this experiment based on a real population, the results are far less dramatic than in Experiment 1. For all the models, the model-dependent SYN estimators SYN−P, LSYN−P and MSYN−D had a substantial bias. The bias was smallest, even though still substantial, for the multilevel-model based estimator MSYN−D. The bias continued to be large even in the large domains. Large bias might be due to the poor fit of the models, even if we used the best models available, and because the inclusion of random effects in the models was quite limited (only a random intercept term was included at the domain level). Accuracy in model-dependent estimators was best again for MSYN−D. As shown in Table 5, there was a slight positive effect of model improvement in MSE.

Table 4
Average Absolute Relative Bias (ARB) (%), Average Relative Root Mean Squared Error (RRMSE) (%) and Average Median Absolute Relative Error (MdARE) (%) of Estimators of the Number of ILO Unemployed in Small and Large Domains (LFS Data)

	Average ARB (%)		Average RRMSE (%)		Average MdARE (%)	
	Expected domain size in sample		Expected domain size in sample		Expected domain size in sample	
	Small (1 – 59)	Large (60+)	Small (1 – 59)	Large (60+)	Small (1 – 59)	Large (60+)
Model-dependent SYN estimators						
SYN−P	36.5	14.2	37.6	16.3	36.6	14.9
LSYN−P	36.4	14.1	37.3	16.2	36.5	14.8
MSYN−D	27.3	9.1	31.8	15.9	29.0	12.1
Model-assisted GREG estimators						
GREG−P	1.2	0.6	46.7	24.0	30.6	16.0
LGREG−P	1.2	0.6	46.8	24.0	30.7	16.0
MGREG−D	1.2	0.6	46.4	24.0	30.6	16.0
Composite estimators						
COMP−D	26.9	8.8	31.8	16.0	28.9	12.1

In model-assisted GREG estimators, the differences in bias and accuracy were small between the multilevel-model assisted MGREG−D estimator and the GREG−P and LGREG−P estimators assisted by population-level fixed

effects models. The fixed-effects linear and logistic models yielded quite similar results, but the multilevel model improved the results slightly, as shown in Table 5.

Table 5
Average Relative Improvement in MSE (%) of Estimators of the Number of ILO Unemployed in Small and Large Domains (LFS Data)

Average relative improvement in MSE (%)	Expected domain size in sample	
	Small (1 – 59)	Large (60+)
MSYN–D versus SYN–P	32.4	1.3
MGREG–D versus GREG–P	0.4	0.2

As measured by the average MdARE, the difference in accuracy between MSYN–D and MGREG–D is small in small domains.

The composite estimates COMP–D were close to the synthetic estimates because the estimated variance of the random intercept was, in most cases, quite small.

7. Summary and Discussion

In the introduction we made a point that, in our opinion, has not been emphasized in earlier literature on domain estimation, namely that the concept "model choice" must be distinguished from the concept "estimator type" when estimation methods are compared. To one and the same choice of model (same mathematical form, same specification of parameters or effects in the model) corresponds one estimator for each of the traditional estimator types discussed in the literature, Synthetic, Generalized Regression, Composite, EBLUP and so on. A first consequence of this is that one cannot make a fair comparison of estimators of different types unless all share the same model choice. Secondly, a change of model, say from a weaker to a stronger model, may have quite different impact on different estimator types. It is this second aspect that is highlighted in this paper.

We have studied the impact of model improvement especially for the Synthetic (SYN) type and Generalized Regression (GREG) type estimators, and found that the impact is very different, and the impact depends heavily of the size of the domain concerned, that is, of the number of sampled units in a domain. Especially in larger domains, the impact of model improvement is very large for SYN type estimators, and modest only for GREG type estimators. The progression is such that a SYN type estimator goes from being highly inaccurate estimator for a weaker model to a much improved estimator for a stronger model. In other words, SYN is highly dependent on the strength of the model. This is not the case for a GREG type estimator. It is slightly more accurate for the stronger model while maintaining a high accuracy for both kinds of models. Its improvement factor is modest compared to a SYN type estimator. We have not carried out our analysis in detail for other estimator types. This is an objective for future research.

The possibilities for efficient estimation for domains and small areas depend on the available statistical infrastructure. As evidenced in many recent papers on small area estimation, one must often start from a set of premises, where the data for model fitting are available not at a unit level, but at some aggregated level (this situation is typical for example in the United Kingdom and in the United States). The background for the methods described in this paper is typical in statistical infrastructures where a good supply of administrative registers exists, with data at the unit level (this holds for example the Scandinavian countries). In such an infrastructure it is often possible to use unit keys, such as personal identification numbers, to merge two or more administrative files at the micro level in building the vector of auxiliary variables. Also, domain membership is often specified for all units in the target population, as assumed in this paper. We can also assume that the collected survey data file can be merged with the auxiliary data file using the unit keys. The situation described above is increasingly found in many countries, for example in several member states of the European Union, where an increasing emphasis is being put on the use of administrative registers for purposes of statistics production.

Technical Appendix

This technical appendix includes the derivation of bias and MSE approximations for GREG and SYN estimators needed for the examination of the effect of model improvement in the case of Simple Random Sampling presented in section 5.

To measure how the accuracy $\hat{Y}_{d\mathrm{GREG}}$ and $\hat{Y}_{d\mathrm{SYN}}$ changes as the model progresses from (4.1) to (4.5), we need to evaluate the variance of each estimator, as well as the bias of $\hat{Y}_{d\mathrm{SYN}}$. By contrast, $\hat{Y}_{d\mathrm{GREG}}$ is nearly unbiased. An obstacle in the analysis of $\hat{Y}_{d\mathrm{GREG}}$ and $\hat{Y}_{d\mathrm{SYN}}$ is their nonlinear form. Therefore we work with the corresponding linearized forms, for which we can easily obtain the bias and the variance. The results are then used to approximate the corresponding characteristics of $\hat{Y}_{d\mathrm{GREG}}$ and $\hat{Y}_{d\mathrm{SYN}}$. Taylor linearization is a standard technique for these types of estimators, as illustrated, for example, in Särndal, Swensson and Wretman (1992), Chapter 6.

Consider first the GREG estimators, GREG–P and GREG–D. Let $\hat{Y}_{d\mathrm{GREG}}$ denote either of those two. With linear approximation, the estimation error (the estimator's deviation from the target parameter Y_d) is

$$\hat{Y}_{d\mathrm{GREG}} - Y_d \approx \sum_s a_k \delta_{dk} E_k - \sum_U \delta_{dk} E_k \qquad \text{(A.1)}$$

where E_k is the population fit residual for k. The difference between GREG–P and GREG–D lies in the residuals E_k. For GREG–P, they are $E_k = E_{Pk}$, where

$E_{Pk} = y_k - \mathbf{x}_k' \mathbf{B}_P$ for $k \in U$, with $\mathbf{B}_P$ given by (4.2). For GREG−D, they are $E_k = E_{dk}$, with $E_{dk} = y_k - \mathbf{x}_k' \mathbf{B}_d$ for $k \in U_d$, $d = 1, ..., D$, with $\mathbf{B}_d$ given by (4.6).

In (A.1), $\sum_s a_k \delta_{dk} E_k$ is the Horvitz-Thompson (HT) estimator for the variable $\delta_{dk} E_k$. Using basic results for the HT estimator we get $E(\hat{Y}_{d\text{GREG}}) - Y_d \approx 0$, that is, $\hat{Y}_{d\text{GREG}}$ is nearly unbiased. It is easy to state the variance for a general design. We need it here for the special case of Simple Random Sampling Without Replacement (SRS). The MSE of $\hat{Y}_{d\text{GREG}}$ equals the variance of $\hat{Y}_{d\text{GREG}}$, to the order of approximation used here.

Next, consider the SYN estimators, SYN−P and SYN−D. Let $\hat{Y}_{d\text{SYN}}$ denote either of those two. After linearization, the estimation error is approximated as

$$\hat{Y}_{d\text{SYN}} - Y_d \approx \sum_s a_k r_{dk} E_k - \sum_U \delta_{dk} E_k \qquad \text{(A.2)}$$

where $E_k = E_{dk}$, $r_{dk} = \delta_{dk}$ for SYN−D, and $E_k = E_{Pk}$, $r_{dk} = R_{dk}$ for SYN−P, with

$$R_{dk} = \left(\sum_{U_d} \mathbf{x}_k\right)\left(\sum_U \frac{\mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1} \frac{\mathbf{x}_k}{c_k}.$$

The term $\sum_s a_k r_{dk} E_k$ in (A.2) is the HT estimator for the variable $r_{dk} E_k$. The quantities R_{dk} vary around a central value at or near the relative domain size, $P_d = N_d / N$. The mean $(1/N)\sum_U R_{dk}$ equals P_d if $\mathbf{x}_k$ contains the constant "1" for every k. From (4.2) we get

$$E(\hat{Y}_{d\text{SYN}}) - Y_d \approx -\sum_{U_d} E_k. \qquad \text{(A.3)}$$

The right hand side of (A.3) is zero for SYN−D, which is therefore nearly unbiased, but is different from zero for SYN−P, which is therefore biased.

For the fixed-effects linear model formulations in section 4.1, we now examine the relative improvement factor (5.1) under SRS with a sampling fraction equal to $f = n / N$.

Consider first the two GREG estimators. We get

$$\begin{aligned} \text{MSE}_T(\hat{Y}_{d\text{GREG}}) &\approx V_T(\hat{Y}_{d\text{GREG}}) \\ &= N^2 \frac{1-f}{n} \frac{1}{N-1} \\ &\quad \sum_U \left\{ \delta_{dk} E_k - \frac{1}{N}\left(\sum_U \delta_{dk} E_k\right)\right\}^2 \end{aligned} \qquad \text{(A.4)}$$

where the index T indicates the approximations derived via the linearized $\hat{Y}_{d\text{GREG}}$, and $E_k = E_{Pk}$ for the P−model and $E_k = E_{dk}$ for the D−model. Developing the square in (A.4) and using $(N_d - 1)/N_d \approx 1$ and $(N_d - 1)/(N-1) \approx N_d / N$ we get

$$\begin{aligned} \text{MSE}_T(\hat{Y}_{d\text{GREG}}) &\approx V_T(\hat{Y}_{d\text{GREG}}) \\ &= N_d^2 \frac{1-f}{n_{d0}} \{S_{EU_d}^2 + (1-P_d)\bar{E}_{U_d}^2\} \end{aligned} \qquad \text{(A.5)}$$

where $n_{d0} = nP_d = n(N_d / N)$ is the expected size of the domain portion of the sample, $s_d = s \cap U_d$, and $S_{EU_d}^2 = (1/(N_d - 1))\sum_{U_d} \{E_k - \bar{E}_{U_d}\}^2$ with $\bar{E}_{U_d} = (1/(N_d))\sum_{U_d} E_k$. If n_{d0} is small, $\hat{Y}_{d\text{GREG}}$ has a poor precision (a high variance), except if the model fits extremely well so that the residual E_k is small for all units in the domain. For GREG−D, $\bar{E}_{U_d} = 0$, so the second term within curly brackets disappears.

Next, consider the two SYN estimators. We get

$$\begin{aligned} \text{MSE}_T(\hat{Y}_{d\text{SYN}}) &= N^2 \frac{1-f}{n} \frac{1}{N-1} \sum_U (r_{dk} E_k)^2 \\ &\quad + N_d^2 \bar{E}_{U_d}^2 \end{aligned} \qquad \text{(A.6)}$$

where r_{dk} and E_k are as specified in (A.2). The first term in (A.6) is the variance; the second is the squared bias obtained from (A.3). The variance term is often very small because the sample size in the denominator is that of the entire sample, not the perhaps much smaller size of the domain part of the sample. The squared bias term is zero for SYN−D, but non-zero, perhaps large, and not tending to zero for SYN−P.

References

Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.

Dempster, A.P., Rubin, D.B. and Tsutakawa, R.K. (1981). Estimation in covariance component models. *Journal of the American Statistical Association*, 76, 341-353.

Djerf, K. (1997). Effects of post-stratification on the estimates of the Finnish Labour Force Survey. *Journal of Official Statistics*, 13, 29-39.

Estevao, V.M., and Särndal, C.-E. (1999). The use of auxiliary information in design-based estimation for domains. *Survey Methodology*, 25, 213-221.

Fay, R.E., and Herriot, R. (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.

Feder, M., Nathan, G. and Pfeffermann, D. (2000). Multilevel modelling of complex survey longitudinal data with time varying random effects. *Survey Methodology*, 26, 53-65.

Fuller, W.A., and Battese, G.E. (1973). Transformations for linear models with nested error structure. *Journal of the American Statistical Association*, 68, 626-632.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association*, 93, 273-282.

Ghosh, M., and Natarajan, K. (1999). Small area estimation: a Bayesian perspective. In *Multivariate Analysis, Design of Experiments, and Survey Sampling* (Ed. S. Ghosh). New York: Marcel Dekker. 69-92.

Goldstein, H. (1995). *Multilevel Statistical Models*. 2nd edition. London: Arnold; New York: John Wiley & Sons, Inc.

Holt, D., and Rao, J.N.K. (1995). Topic 3: Small area estimation. *Bulletin of the International Statistical Institute* 50th session, 56 (book 4), 1648-1650.

Holt, D., Smith, T.M.F. and Tomberlin, T.J. (1979). A model-based approach to estimation for small subgroups of population. *Journal of the American Statistical Association*, 74, 405-410.

Lehtonen, R., and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, 24, 51-55.

Lehtonen, R., and Veijanen, A. (1999). Domain estimation with logistic generalized regression and related estimators. Proceedings, IASS Satellite Conference on Small Area Estimation, Riga, August 1999. Riga: Latvian Council of Science, 121-128.

McCullogh, C.E., and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. New York: John Wiley & Sons, Inc.

Marker, D. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.

Moura, F.A.S., and Holt, D. (1999). Small area estimation using multilevel models. *Survey Methodology*, 25, 73-80.

National Center for Heath Statistics (1968). Synthetic State Estimates of Disability. PHS publication no. 1959. Washington, DC: Public Health Service, US Government Printing Office.

National Research Council (1980). Panel on Small-Area Estimates of Population and Income. Estimating Population and Income of Small Areas. Washington, DC: National Academy Press.

Pfeffermann, D. (1999). Small area estimation - big developments. Proceedings, IASS Satellite Conference on Small Area Estimation, Riga, August 1999. Riga: Latvian Council of Science, 129-145.

Prasad, N.G.N., and Rao, J.N.K. (1999). On robust small area estimation using a simple random effects model. *Survey Methodology*, 25, 67-72.

Rao, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25, 175-186.

Rao, J.N.K. (2003). *Small Area Estimation*. Hoboken, NJ: John Wiley & Sons, Inc.

Särndal, C.-E., Swensson, B. and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Singh, A.C., Stukel, D.M. and Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society*, *B*, 60, 377-396.

You, Y., and Rao, J.N.K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology*, 26, 173-181.