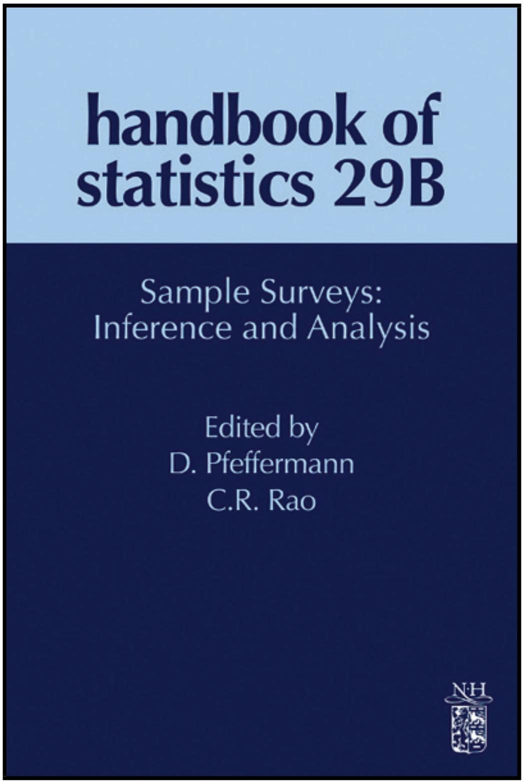
handbook of statistics 29B
Sample Surveys: Inference and Analysis
Edited by
D. Pfeffermann
C.R. Rao
NH

From Risto Lehtonen and Ari Veijanen, Design-based Methods of Estimation for Domains and Small Areas.
In: D. Pfeffermann and C. R. Rao, editors, Handbook of Statistics – Sample Surveys:
Inference and Analysis, Volume 29B.
The Netherlands: North-Holland, 2009, pp. 219–250.
ISBN: 978-0-444-53438-5

North-Holland

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(09)00231-4

31

Design-based Methods of Estimation for Domains and Small Areas

Risto Lehtonen and Ari Veijanen

1. Introduction

This chapter is devoted to the estimation for population subgroups or domains. Regional areas constructed by administrative criteria, such as county or municipality, are typical *domains of study* (Yates, 1949), also called *domains of interest*. *Estimation for domains*, or *domain estimation* for short, refers to the estimation of population quantities, such as totals or means, for the desired population subgroups. Domain estimation will be examined in the context of *design-based estimation*. Design-based methods for domain estimation are frequently used in many areas of empirical research and official statistics production.

Design-based estimation for a finite population quantity refers to an estimation approach where the randomness is introduced by the sampling design. Thus, the approach also is called *randomization approach*. In design-based estimation, it is emphasized that estimators should be design consistent and, preferably, essentially (or nearly) design unbiased at least in medium-sized samples.

Some early milestones of design-based estimation for domains are Yates (1953, 1960) and Durbin (1958). Hartley (1959) introduced the so-called domain-specific variables for domain estimation with standard design-based estimators of population quantities. This technique has appeared fruitful for example in software development for domain estimation.

We focus on the estimation of *population totals* for domains. Totals are chosen because of their fundamental role in survey sampling and because more complex parameters can often be expressed as functions of totals. The estimation of *ratios* and *quantiles*, such as median, is also discussed. The availability of high-quality *auxiliary information* is crucial for reliable estimation for domains. The reason for incorporating auxiliary data in a domain estimation procedure is obvious: improved accuracy is attained if strong auxiliary data are available for domain estimation.

Different types of auxiliary data can be used in design-based estimation for domains. The available auxiliary data can be aggregated at the population level, at the domain level, or at an intermediate level. Aggregates are often taken from reliable auxiliary sources such as population census or other official statistics; this case is common, for example in North America. If the auxiliary data are included in a sampling frame, as is the case in many European countries, notably in Scandinavia, the necessary auxiliary totals can be aggregated at the desired level from unit-level data sources.

Calibration techniques and *model-assisted methods* using aggregated auxiliary data offer efficient tools for design-based domain estimation. Calibration is discussed, for example, in Deville and Särndal (1992) and Kott (2003). Särndal (2007) provides a comprehensive treatment of the calibration approach in survey theory and practice. An overview on calibration weighting is given in Chapter 25. Calibration methods were developed for domain estimation in Estevao and Särndal (1999, 2006). The proposed approach to calibration is sometimes called linear or *model-free calibration*. Model-assisted methods using *generalized regression* (GREG) *estimators* were extensively discussed in Särndal et al. (1992). GREG estimation was introduced for domain estimation in Särndal (1981, 1984), Hidiroglou and Särndal (1985), and Särndal and Hidiroglou (1989) and were developed further (including computational tools) in Estevao et al. (1995). We elaborate to some extent these developments; it will appear that the level at which the auxiliary data are used is crucial: efficiency tends to improve when the aggregation level comes close to the domain level when compared to the use of higher-level aggregates.

A statistician also can be in a favorable position to use unit-level auxiliary data for domain estimation. These data are incorporated in the estimation procedure by unit-level statistical models. We illustrate various members of the family of GREG estimators for these cases. For this purpose, we assume that register data (such as population census register, business register, different administrative registers) are available as frame populations and sources of auxiliary data, and the registers contain unique identification keys that can be used in merging at microlevel data from registers and sample surveys. Known domain membership for all population elements is often assumed. Many countries, both in Europe and elsewhere, are progressing in the development of reliable population and business registers that can be accessed for statistical purposes. Obviously, access to micromerged register and survey data provides great flexibility for domain estimation. In GREG estimation, this view has been adopted, for example, in Lehtonen and Veijanen (1998), Särndal (2001), Lehtonen et al. (2003, 2005), and Hidiroglou and Patak (2004). Wu and Sitter (2001a) use unit-level auxiliary information in their *model calibration* method.

Design-consistent estimation for domains contrasts with *model-dependent* estimators, which can have desirable properties under the model but whose design bias does not necessarily tend to zero with increasing sample size (Hansen et al., 1978, 1983; Lehtonen et al., 2003; Särndal, 1984). Design-consistent domain estimators also have been proposed in the context of *model-based* estimation. Model-based and model-dependent methods falling under the headline of *small-area estimation* may be required for the smallest domains (with a small sample size in a domain), where design-based estimators often fail. The methods include a variety of model-based techniques such as synthetic and composite estimators, empirical best linear unbiased predictor (EBLUP) type estimators and various Bayesian techniques. The monograph by Rao (2003a) provides a comprehensive treatment of model-based small-area estimation. Model-based small-area estimation is discussed in Chapter 32.

In design-based estimation, the existence of a model is not necessarily recognized. For example in model-free calibration, an explicit model is not present but exists in the model calibration method. An assisting or "working" model is postulated in model-assisted estimation. In GREG estimation, the main goal is to obtain favorable design-based properties, such as small design bias. These design-based properties should hold even when the model is misspecified. If our model fits well, decreased design variance is expected for a GREG estimator. Thus, a model is used as an assisting tool in constructing the estimator, which is then modified to meet the desired design-based properties. For example, a GREG estimator for a domain total is often constructed by adding a bias correction term to the sum of fitted values calculated over the population domain. The bias correction term is obtained as a weighted sum of the sample residuals over the domain.

In this chapter, we do not address design-based techniques for nonresponse adjustment (see Chapter 8). Calibration approach to nonresponse treatment is discussed in Särndal and Lundström (2005). Additional topics that are not covered include informative sampling in the context of domain estimation (e.g., Pfeffermann and Sverchkov, 2007) and estimation for domains in the presence of outliers (see Chapter 11).

This chapter is organized as follows. Theoretical framework, terminology, and notation are introduced in Section 2. Section 3 discusses direct estimation for domains by the Horvitz–Thompson (HT) estimator, calibration and GREG estimators. In these cases, domains are often considered as strata in the sampling design. We extend in Section 4 our discussion to more general estimator types and domain structures that are often encountered in practice. GREG estimators for domains are discussed extensively; we also address composite estimation from a design-based perspective. In all these cases, auxiliary information is needed at an aggregated level. Extensions are discussed in Section 5, where a number of empirical examples based on simulation experiments are presented. In these cases, access to unit-level auxiliary data is assumed. Section 6 summarizes some properties of selected software products that can be used for design-based domain estimation.

2. Theoretical framework, terminology, and notation

2.1. Design-based inference at the population level

Let us consider a collection of random variables $(Y_1, Y_2, \ldots, Y_k, \ldots, Y_N)$ with unknown values $(y_1, y_2, \ldots, y_k, \ldots, y_N)$ of a variable of interest y in a *fixed* and *finite population* $U = \{1, 2, \ldots, k, \ldots, N\}$, where k refers to the label of population element. The fixed population is said to be generated from a *superpopulation*. For practical purposes, we are interested in one particular realized population U with $(y_1, y_2, \ldots, y_N)$, not in the more general properties of the model explaining how the population evolved. This is important especially in national statistical agencies, which attempt to describe the current state of the population of a country.

In the design-based approach, the values of the variable of interest are regarded as fixed but unknown quantities. The only source of randomness is the sampling design, and our conclusions should apply to hypothetical repeated sampling from the fixed population.

In estimation for the whole population, we are mainly interested in the total $t = \sum_{k\in U} y_k$ or mean $\bar{y} = \sum_{k\in U} y_k/N$ of the variable y. Notation $\sum_{k\in U}$ refers to summation over all population units $k \in U$. In practice, the values y_k of y are observed in an n element sample $s \subset U$, which is drawn at random by a sampling design giving probability $p(s)$ to each sample s. The sampling design can be *complex* involving stratification and clustering and several sampling stages.

The design expectation of an estimator $\hat{t}$ of population total t is determined by the probabilities $p(s)$: let $\hat{t}(s)$ denote the value of estimator that depends on y observed in s. Then the expectation is $E(\hat{t}) = \sum_s p(s)\hat{t}(s)$. A *design unbiased* estimator has $E(\hat{t}) = t$. *Design variance* is defined as $\mathrm{Var}(\hat{t}) = \sum_s p(s) \left(\hat{t}(s) - E(\hat{t})\right)^2$. An estimator of design variance is denoted by $\hat{V}(\hat{t})$.

An estimator is *design consistent* if its design bias and variance tend to zero as the sample size increases. An estimator is *nearly design unbiased* if its bias ratio (bias divided by standard deviation) approaches zero with order $O(n^{-1/2})$ when the total sample size n tends to infinity (Estevao and Särndal, 2004). For a nearly design unbiased estimator, the design bias is, under mild conditions, an asymptotically insignificant contribution to the estimator's mean squared error (MSE) (Särndal, 2007, p. 99).

Variance estimators are derived in two steps. First, the theoretical design-based variance $\mathrm{Var}(\hat{t})$ (or its approximation if the theoretical design variance is intractable) is derived. Second, the derived quantity is estimated by a design unbiased or design-consistent estimator $\hat{V}(\hat{t})$.

When the estimator is a weighted sum of observations over sample, it is practical to derive expectation and variance using *inclusion probabilities*. An observation k is included in the sample with probability $\pi_k = P\{k \in s\}$. The inverse probabilities are called *design weights* $a_k = 1/\pi_k$. A useful tool is a sample membership indicator $I_k = I\{k \in s\}$ with value 1 if k is in the sample and 0 otherwise, $E(I_k) = \pi_k$. In variances, we have to consider inclusion of pairs of observations: the probability of including both k and $l (k \neq l)$ is $\pi_{kl} = E(I_k I_l)$ with inverse $a_{kl} = 1/\pi_{kl}$, and $a_{kl} = a_k$ when $k = l$. The covariance of I_k and I_l is $\mathrm{Cov}(I_k, I_l) = \pi_{kl} - \pi_k \pi_l$; this quantity is needed in constructing design variances and their estimators, especially for without-replacement type designs.

2.2. *Basic features of design-based inference for domains*

2.2.1. *Planned and unplanned domain structures*

In domain estimation, we are mainly interested in totals or averages of a variable of interest y over D nonoverlapping domains $U_d \subset U$, $d = 1, 2, \ldots, D$, with possibly known domain sizes N_d. As an example, consider the population of a country divided into D domains by regional classification, with N_d households in domain U_d, and the aim is to estimate statistics on household poverty for the regional areas. A domain total is $t_d = t_{dy} = \sum_{k \in U_d} y_k$, where y_k refers to measurement for household k, and domain mean is $\bar{y}_d = t_d/N_d$, $d = 1, \ldots, D$.

Corresponding to population domains, the sample s is divided into subsamples s_d, $d = 1, \ldots, D$. Sampling design may be based on knowledge of domain membership of units in population. If the sampling design is stratified, domains being the strata, the domains are called *planned* (Singh et al., 1994) or *primary domains* (Hidiroglou and Patak, 2004); sometimes also *design domains* (Kish, 1980) or *identified domains* (Särndal, 2007). For planned domain structures, the population domains U_d can be regarded as separate subpopulations. Therefore, standard population estimators are applicable as such. The domain size N_d in every domain U_d is often assumed known and the sample size n_d in domain sample $s_d \subset U_d$ is fixed in advance. Stratified sampling in connection to a suitable allocation scheme such as optimal (Neyman) or power (Bankier) allocation is advisable in practical applications to obtain control over domain sample sizes (e.g., Lehtonen and Pahkinen, 2004). Singh et al. (1994) describe allocation strategies to attain reasonable accuracy for small domains, still retaining good accuracy for large domains. Falorsi et al. (2006) propose sample balancing and coordination techniques for cases with a large number of different stratification structures to be addressed in domain estimation. If the domain membership is not incorporated into the sampling design, the sizes n_{s_d} of domain samples $s_d = s \cap U_d$ will be random. The domains are then called *unplanned* or *secondary domains*. Unplanned domain structures typically cut across design strata. The property of random domain sample sizes introduces an increase in the variance of domain estimators. In addition, extremely small number (even zero) of sample elements in a domain can be realized if the domain size in the population is small. Unplanned domain structures are commonly encountered in practice because it is impossible to include all relevant domain structures into the sampling design of a given survey.

2.2.2. *Extended domain variables of interest*

A general tool for domain estimation is the *extended domain variable of interest* y_d defined as $y_{dk} = y_k$ for $k \in U_d$ and $y_{dk} = 0$ for $k \notin U_d$ (Hartley, 1959). In other words, $y_{dk} = I\{k \in U_d\}y_k$. Because $t_d = \sum_{k \in U_d} y_k = \sum_{k \in U} y_{dk}$, we can estimate the domain total of y by estimating the population total of y_{dk} (e.g., Estevao et al., 1995; Estevao and Särndal, 1999; Hidiroglou and Patak, 2004). Consequently, any population total or mean estimator applied to y_{dk} is usable as a corresponding domain estimator. Extended domain variables are useful for estimation for unplanned domains because the contribution of extra variance caused by random domain sample sizes can be easily incorporated in variance expressions. The technique of extended domain variables allows building of generally applicable software for domain estimation and is implemented, for example, in survey sampling oriented SAS procedures and the GES software of Statistics Canada (Estevao et al., 1995).

Extended domain variables can be incorporated in a model-assisted estimation procedure. However, a model fitted to the whole sample is not always going to fit well because most of the y_{dk} are zeroes. But when using extended domain variables, the main interest is not necessarily in the goodness of fit; the primary objective is to attain a single set of weights for all domains. Moreover, the estimates are additive: their sum over the domains equals the estimate for the whole population (Estevao et al., 1995). This can be considered as a benefit of practical importance, especially for routine official statistics production. On the other hand, possible efficiency gains might not be attained and therefore, we usually attempt to derive estimators using the original y_k values.

2.2.3. *Direct and indirect estimators*

It is advisable to separate direct and indirect estimators for domains. A *direct* estimator uses values of the variable of interest only from the time period of interest and only from units in the domain of interest (Federal Committee on Statistical Methodology, 1993). A HT type estimator $\hat{t}_d = \sum_{k \in s_d} y_k/\pi_k$ provides a simple example of direct estimator. In model-assisted estimation, direct estimators are constructed by using models fitted separately in each domain; an example is a model $Y_k = \mathbf{x}'_k \boldsymbol{\beta}_d + \varepsilon_k$, $k \in U_d$, with domain-specific auxiliary x-data and a vector of regression coefficients

$\boldsymbol{\beta}_d$, $d = 1, \ldots, D$. A direct domain estimator can still incorporate auxiliary data outside the domain of interest. This is relevant if accurate population data about the auxiliary x-variables are only available at a higher aggregate level.

An *indirect* domain estimator uses values of the variable of interest from a domain and/or time period other than the domain and time period of interest (Federal Committee on Statistical Methodology, 1993). For example, if a linear model $Y_k = \mathbf{x}'_k\boldsymbol{\beta} + \varepsilon_k$, $k \in U$, with a common vector $\boldsymbol{\beta}$ is used as an assisting model, the resulting domain estimator will be indirect. In general, indirect estimators are attempting to "borrow strength" from other domains and/or in a temporal dimension. The concept of "borrowing strength" is often used in model-based small-area estimation (e.g., Rao, 2003a). Indirect model-assisted estimators for domains are discussed in the literature (e.g., Estevao and Särndal, 1999; Hidiroglou and Patak, 2004; Lehtonen et al., 2003, 2005). Estevao and Särndal (2004) have argued in favor of direct estimators in the context of design-based estimation for domains.

2.2.4. *Conditional design-based inference for domains*

For unplanned domain structures, observed domain sample sizes can be taken into account in estimation and in theory. We are interested in the average properties of estimators in samples with observed domain sample sizes $\mathbf{n} = (n_1, n_2, \ldots, n_d, \ldots n_D)'$. In conditional design-based inference for domains (Falorsi et al., 2000; Hidiroglou and Patak, 2004; Särndal and Hidiroglou, 1989) given $\mathbf{n}$, the hypothetical repeated sampling yields only samples s with $\mathbf{n}(s) = \mathbf{n}$. This subset of samples, $S_{\mathbf{n}} = \{s^* \subset U : \mathbf{n}(s^*) = \mathbf{n}\}$, is based on observed information, so it has been considered more relevant than the set of all possible samples. By using conditional probabilities $p_c(s) = p(s)/P\{\mathbf{n}(s^*) = \mathbf{n}\}$, if $\mathbf{n}(s) = \mathbf{n}$, and $p_c(s) = 0$ otherwise, the conditional expectation of an estimator is defined as $E(\hat{t}_d|\mathbf{n}(s) = \mathbf{n}) = \sum_{s:\mathbf{n}(s)=\mathbf{n}} p_c(s)\hat{t}_d(s)$. The conditional MSE and variance are defined in the same way.

We prefer conditionally unbiased estimators to conditionally biased ones. We do not encounter estimators that are conditionally unbiased but unconditionally biased because the unconditional expectation is an average over conditional expectations. The conditional approach may also result in changes in a domain estimation procedure. For example, Falorsi et al. (2000) introduced a HT type estimator and a ratio estimator incorporating conditional inclusion probabilities. Park and Fuller (2005) used conditional inclusion probabilities for a calibrated GREG estimator.

The estimator of the conditional variance is, in general, different from the estimator of the unconditional variance. Conditional variance estimate yields a conditional confidence interval. In repeated sampling from the subset $S_{\mathbf{n}}$, the conventional t-based conditional confidence interval covers the true value approximately at a given rate if the estimator is approximately normally distributed. Because this holds for all values of $\mathbf{n}$, the conditional confidence interval is also an unconditional confidence interval with the same coverage rate. If the model is only approximately correct, a model-assisted method does not always yield conditionally valid inference. It can be argued (Rao, 1997) that model-assisted approach should be restricted to methods with good conditional properties. Conditional inference has been based on other properties besides domain sizes; there are examples of conditioning on strata sample sizes (Holt and Smith, 1979) and on HT estimates of the auxiliary variables (Montanari and Ranalli, 2002; Rao, 1985).

2.2.5. Design-based properties of domain estimators

Known design-based properties related to bias and accuracy of model-assisted estimators are summarized in Table 1. For comparison, design-based properties of corresponding model-dependent estimators are also included in the table. Model-assisted estimators such as GREG are design consistent or nearly design unbiased by definition, but their variance can become large in domains where the sample size is small. Model-dependent estimators such as synthetic and EBLUP estimators are design biased: the bias can be large for domains where the model does not fit well. The variance of a model-dependent estimator can be small even for small domains, but the accuracy can be poor if the squared bias dominates the MSE, as shown, for example, by Lehtonen et al. (2003, 2005). For a model-dependent estimator, the dominance of the bias component together with a small variance can cause poor coverage rates and invalid design-based confidence intervals. For design-based model-assisted estimators, on the other hand, valid confidence intervals can be constructed. Typically, model-assisted estimators are used for major or not-so-small domains, and model-dependent estimators are used for small domains where model-assisted estimators can fail.

Table 1 indicates that small domains present problems in the design-based approach. Purcell and Kish (1980) call domain a minidomain when $N_d/N < 1\%$. In such small domains, especially, direct estimators can have large variance. Small domains are the main reason to prefer indirect model-based estimators to design-based estimators (Rao, 2005). By proper planning of the sampling strategy, it is possible to decrease the variance of a design-based estimator in the small domains. Singh et al. (1994) and Marker (2001) give examples of such strategies.

In practice, there are two main approaches to design-based estimation for domains: direct estimators that are usually applied for planned domain structures and indirect estimators whose natural applications are for unplanned domains. The two main approaches are discussed in Sections 3 and 4, respectively.

Table 1
Design-based properties of model-assisted and model-dependent estimators for domains and small areas

	Design-based model-assisted methods	Model-dependent methods
	GREG and calibration estimators	Synthetic and EBLUP estimators
Bias	Design unbiased (approximately) by the construction principle	Design biased Bias does not necessarily approach zero with increasing domain sample size
Precision (Variance)	Variance may be large for small domains Variance tends to decrease with increasing domain sample size	Variance can be small even for small domains Variance tends to decrease with increasing domain sample size
Accuracy (MSE)	MSE = Variance (or nearly so)	MSE = Variance + squared bias Accuracy can be poor if the bias is substantial
Confidence intervals	Valid design-based intervals can be constructed	Valid design-based intervals not necessarily obtained

3. Direct estimators for domain estimation

The HT type estimator does not incorporate auxiliary information. GREG estimation is assisted by a model fitted at the domain level and uses auxiliary data from the domain. Calibration incorporates auxiliary data from the domain of interest or from a higher-level aggregate. All these estimators are direct because the y-values are taken from the domain of interest. When domain membership is known for all population elements, domain sizes N_d are also known.

3.1. Horvitz–Thompson estimator

The basic design-based direct estimator of the domain total t_d is the HT estimator, also known as the Narain-Horvitz-Thompson (NHT) and the *expansion estimator*:

$$\hat{t}_{d\mathrm{HT}} = \sum_{k \in U_d} I_k y_k / \pi_k = \sum_{k \in s_d} y_k / \pi_k = \sum_{k \in s_d} a_k y_k \tag{1}$$

(Horvitz and Thompson, 1952; Narain, 1951; notation as in Section 2.1). HT estimates of domain totals are additive: they sum up to the HT estimator $\hat{t}_{\mathrm{HT}} = \sum_{k \in s} a_k y_k$ of the population total. As $E(I_k) = \pi_k$, the HT estimator is design unbiased for t_d. Under mild conditions on the π_k, the corresponding mean estimator $\hat{t}_{d\mathrm{HT}}/N_d$ is also design consistent (Isaki and Fuller, 1982). The estimator $\hat{t}_{d\mathrm{HT}}$ has design variance

$$\begin{aligned}\mathrm{Var}\big(\hat{t}_{d\mathrm{HT}}\big) &= E\left(\sum_{k \in U_d} \frac{I_k - \pi_k}{\pi_k} y_k\right)^2 = \sum_{k \in U_d} \sum_{l \in U_d} E(I_k - \pi_k)(I_l - \pi_l)\,\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} \\ &= \sum_{k \in U_d} \sum_{l \in U_d} (\pi_{kl} - \pi_k \pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} = \sum_{k \in U_d} \sum_{l \in U_d} (a_k a_l / a_{kl} - 1) y_k y_l. \end{aligned} \tag{2}$$

From $a_{kl} E(I_k I_l) = 1$, we see that an unbiased estimator for the design variance is

$$\hat{V}\big(\hat{t}_{d\mathrm{HT}}\big) = \sum_{k \in U_d} \sum_{l \in U_d} a_{kl} I_k I_l (a_k a_l / a_{kl} - 1) y_k y_l = \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl}) y_k y_l. \tag{3}$$

An alternative Sen–Yates–Grundy formula for fixed sample size designs is (Sen, 1953; Yates, 1953):

$$\begin{aligned}\hat{V}\big(\hat{t}_{d\mathrm{HT}}\big) &= -\sum_{k \in s_d} \sum_{l<k; l \in s_d} a_{kl}(\pi_{kl} - \pi_k \pi_l)(a_k y_k - a_l y_l)^2 \\ &= \sum_{k \in s_d} \sum_{l<k; l \in s_d} (a_{kl}/a_k a_l - 1)(a_k y_k - a_l y_l)^2. \end{aligned}$$

These variance estimators are impractical because they contain second-order inclusion probabilities π_{kl} whose computation is often laborious for practical purposes. Hájek (1964) and Berger (2004, 2005b) proposed approximations to π_{kl}. Särndal (1996) developed efficient strategies with simple variance estimators under fixed sample size probability proportional-to-size (πPS) schemes, including a combination of Poisson sampling or stratified simple random sampling without replacement (SRSWOR) with

GREG estimation. Berger and Skinner (2005) proposed a jackknife variance estimator and Kott (2006a) introduced a delete-a-group jackknife variance estimator for πPS designs. The SAS procedure SURVEYSELECT is able to compute π_{kl} under certain unequal probability without-replacement sampling designs. Some software products can incorporate the π_{kl} into variance estimation procedures; an example is the SUDAAN software. The SAS macro CLAN includes the Sen–Yates–Grundy formula. Such estimators are discussed in Chapter 2.

Many πPS designs allow using of Hájek approximation (Berger, 2004, 2005b; Hájek, 1964) of second-order inclusion probabilities by $\pi_{kl} \approx \pi_k \pi_l \left[1-(1-\pi_k)(1-\pi_l)m_d^{-1}\right]$ for $k \neq l$, where $m_d = \sum_{i \in U_d} \pi_i(1-\pi_i)$. The approximation is used in a simple variance estimator $\hat{V}\left(\hat{t}_{d\mathrm{HT}}\right) = \sum_{k \in s_d} c_k e_k^2$, where $c_i = n_d(n_d-1)^{-1}(1-\pi_i)$ and $e_k = a_k y_k - \left(\sum_{i \in s_d} c_i\right)^{-1} \sum_{i \in s_d} c_i a_i y_i$.

For unequal probability sampling designs, the variance of the ordinary HT estimator has been approximated under a with-replacement (WR) assumption, leading to Hansen–Hurwitz (1943) type variance estimator (Lehtonen and Pahkinen, 2004, p. 228, and SAS procedure SURVEYMEANS) given by

$$\hat{V}_A(\hat{t}_{d\mathrm{HT}}) = \frac{1}{n_d(n_d-1)} \sum_{k \in s_d} (n_d a_k y_k - \hat{t}_{d\mathrm{HT}})^2. \tag{4}$$

For unplanned domains, the variance estimator for HT should account for random domain sizes. An approximate variance estimator applied, for example, in SAS procedure SURVEYMEANS contains extended domain variables y_{dk}:

$$\hat{V}_U(\hat{t}_{d\mathrm{HT}}) = \frac{n}{n-1} \sum_{k \in s} (a_k y_{dk} - \hat{t}_d/n)^2, \tag{5}$$

where n is the total sample size. Under SRSWOR, an alternative to (5) is

$$\hat{V}_{\mathrm{srswor}}(\hat{t}_{d\mathrm{HT}}) = N^2\left(1-\frac{n}{N}\right)\left(\frac{1}{n}\right) p_d \hat{s}_{dy}^2 \left(1+\frac{q_d}{\mathrm{c.v}_{dy}^2}\right),$$

where $p_d = n_{s_d}/n$, $q_d = 1 - p_d$, variance estimator is, $\hat{s}_{dy}^2 = \sum_{k \in s_d}(y_k - \bar{y}_d)^2/(n_{s_d} - 1)$, and estimated coefficient of variation is $\mathrm{c.v}_{dy} = \hat{s}_{dy}/\bar{y}_d$ for $\bar{y}_d = \sum_{k \in s_d} y_k / n_{s_d}$.

The HT estimator can be regarded as a model-dependent estimator under a model $Y_k = \beta\pi_k + \pi_k\varepsilon_k$ (Zheng and Little, 2003). HT is nearly optimal estimator among weighted sums of Y values when Y depends on scalar x as $E(Y_k) = \beta x_k$, the variance of errors is proportional to x_k^2, and the sampling design assigns π_k proportional to x_k. On the other hand, HT is very inefficient when the intercept of the model is far from zero. Disastrous results are possible in HT estimation, as the famous example of Basu (1971) shows (e.g., citation in Little, 2004).

If the domain size N_d is known, we expect better results with a "Hájek" type direct estimator $\hat{t}_{dH(N)} = N_d \hat{\bar{y}}_d$ (e.g., Hidiroglou and Patak, 2004; Särndal et al., 1992, p. 391) derived from the domain mean $\hat{\bar{y}}_d = \sum_{k \in s_d} a_k y_k / \hat{N}_d$ with $\hat{N}_d = \sum_{k \in s_d} a_k$. This is a special case of ratio estimation (Section 4.3.1). The variance of $\hat{t}_{dH(N)}$ is estimated by

$$\hat{V}(\hat{t}_{dH(N)}) = \left(\frac{N_d}{\hat{N}_d}\right)^2 \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl})(y_k - \hat{\bar{y}}_d)(y_l - \hat{\bar{y}}_d). \tag{6}$$

3.2. Population fit regression estimator

The population fit regression estimator is a theoretical tool used in approximating real-world estimators. We first consider *difference estimators* (Särndal, 1980; Särndal et al., 1992, p. 221). If known values y_k^0 are close to y_k, we write the estimable population total as

$$t = \sum_{k \in U} y_k = \sum_{k \in U} y_k^0 + \sum_{k \in U} (y_k - y_k^0).$$

A difference estimator is defined by estimating the second sum using HT:

$$\hat{t}_{\mathrm{DIFF}} = \sum_{k \in U} y_k^0 + \sum_{k \in s} a_k (y_k - y_k^0).$$

As the y_k^0 are constants, $\hat{t}_{\mathrm{DIFF}}$ is unbiased for t.

Consider a regression superpopulation model $Y_k = \mathbf{x}'_k \boldsymbol{\beta} + \varepsilon_k$, where $\mathbf{x}_k = (1, x_{1k}, \ldots, x_{Jk})'$ is the vector of auxiliary x-variables, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_J)'$ is the vector of regression coefficients, and ε_k are the residuals with variances $\sigma_k^2 = \mathrm{Var}(\varepsilon_k)$. Hypothetically, we can fit the model to the population by calculating generalized least squares (GLS) estimator $\mathbf{B} = \hat{\boldsymbol{\beta}}$ as

$$\mathbf{B} = \left(\sum_{k \in U} \frac{\mathbf{x}_k \mathbf{x}'_k}{\sigma_k^2} \right)^{-1} \left(\sum_{k \in U} \frac{\mathbf{x}_k y_k}{\sigma_k^2} \right).$$

In practice, the error variance $\mathrm{Var}(\varepsilon_k) = \sigma_k^2$ can often be assumed constant, $\sigma_k^2 = \sigma^2$, and then it cancels out. When the variance varies between observations, the σ_k^2 should be included in the estimators. Straightforward cases are known σ_k^2 or an assumption that the variances differ by known constants c_k such that $\sigma_k^2 = c_k \sigma^2$. A special case is when $c_k = 1$ for all $k \in U$. For more details on the treatment of σ_k^2, see, for example, Särndal et al. (1992, p. 229 and Chapter 7).

A difference estimator with fitted values $\hat{y}_k^0 = \mathbf{x}'_k \mathbf{B}$ defines the *population fit regression estimator*,

$$\hat{t}_{\mathrm{REG}} = \sum_{k \in U} \hat{y}_k^0 + \sum_{k \in s} a_k (y_k - \hat{y}_k^0).$$

If an estimator $\hat{t}$ can be well approximated by $\hat{t}_{\mathrm{REG}}$, then $\mathrm{Var}(\hat{t})$ can be estimated by a sample-based estimator of

$$\mathrm{Var}(\hat{t}_{\mathrm{REG}}) = \mathrm{Var}\left(\sum_{k \in s} a_k E_k \right) = \sum_{k \in U} \sum_{l \in U} (a_k a_l / a_{kl} - 1) E_k E_l,$$

where $E_k = y_k - \hat{y}_k^0$ are the population fit residuals. To estimate $\mathrm{Var}(\hat{t}_{\mathrm{REG}})$ from sample, we replace the E_k by corresponding sample residuals $e_k = y_k - \mathbf{x}'_k \hat{\mathbf{B}}$. If $\hat{\mathbf{B}}$ is nearly unbiased for $\mathbf{B}$, we can verify using $E(a_{kl} I_k I_l) = 1$ that a nearly unbiased estimator for $\mathrm{Var}(\hat{t}_{\mathrm{REG}})$ is

$$\hat{V}(\hat{t}_{\mathrm{REG}}) = \sum_{k \in s} \sum_{l \in s} (a_k a_l - a_{kl}) e_k e_l. \tag{7}$$

One approach to estimate $\mathbf{B}$ is to plug in HT estimators of both of its sum components. When σ_k^2 is constant, we use a weighted least squares (WLS) estimator

$$\hat{\mathbf{B}} = \left(\sum_{k\in s} a_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum_{k\in s} a_k \mathbf{x}_k y_k\right).$$

This estimator is only approximately unbiased due to its nonlinearity. Another approach is to consider the population maximum likelihood (ML) estimator maximizing $f(\boldsymbol{\beta}) = -\sum_{k\in U} \left(y_k - \mathbf{x}'_k\boldsymbol{\beta}\right)^2/\sigma^2$. As only the sample is available, we use an estimated log-likelihood, the so-called *pseudolikelihood*, instead (Binder, 1983; Godambe and Thompson, 1986a; Nordberg, 1989). The function $f(\boldsymbol{\beta})$ is estimated by an unbiased HT type estimator $\hat{f}(\boldsymbol{\beta}) = -\sum_{k\in s} a_k \left(y_k - \mathbf{x}'_k\boldsymbol{\beta}\right)^2/\sigma^2$. This function is maximized by $\hat{\mathbf{B}}$. Robust alternatives are presented in Beaumont and Alavi (2004).

Särndal et al. (1992) and Estevao and Särndal (2006) have approximated GREG and calibration estimators (Sections 3.3 and 3.4) by Taylor linearization yielding a population fit regression estimator. Because many approximations are involved, the resulting variance estimators are at least slightly biased.

3.3. *GREG estimators*

The GREG estimator is a sample-based substitute for the population fit regression estimator (Section 3.2). A direct type GREG estimator of domain total t_d is assisted by a regression model $Y_k = \mathbf{x}'_k\boldsymbol{\beta}_d + \varepsilon_k$, $\mathrm{Var}(\varepsilon_k) = \sigma_k^2$. Assuming constant error variance σ_k^2, the domain-specific parameter $\mathbf{B}_d$ of the population fit defined for U_d is estimated as in Section 3.2 by

$$\hat{\mathbf{B}}_d = \left(\sum_{k\in s_d} a_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum_{k\in s_d} a_k \mathbf{x}_k y_k\right),$$

and the fitted values $\hat{y}_k = \mathbf{x}'_k\hat{\mathbf{B}}_d$ and residuals $e_k = y_k - \hat{y}_k$ are incorporated into the GREG estimator

$$\hat{t}_{d\mathrm{GREG}} = \sum_{k\in U_d} \hat{y}_k + \sum_{k\in s_d} a_k(y_k - \hat{y}_k) = \sum_{k\in U_d} \hat{y}_k + \sum_{k\in s_d} a_k e_k \qquad (8)$$

(Särndal, 1980; Särndal et al., 1992). The first part in $\hat{t}_{d\mathrm{GREG}}$, the population sum of fitted values over the domain, is sometimes called a synthetic estimator (Särndal, 1984). When compared with direct GREG, it may have smaller variance but possibly large design bias. The weighted sum of residuals tends to correct for the design bias. In some cases, however, the weighted sum of the residual terms is zero. This happens when the model contains an intercept.

Rearranging the terms of GREG we obtain the traditional regression estimator

$$\hat{t}_{d\mathrm{GREG}} = \hat{t}_{d\mathrm{HT}} + (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx})'\hat{\mathbf{B}}_d,$$

where $\mathbf{t}_{dx} = \sum_{k\in U_d} \mathbf{x}_k = \left(N_d, \sum_{k\in U_d} \mathbf{x}_{1k}, \ldots, \sum_{k\in U_d} \mathbf{x}_{Jk}\right)'$ and $\hat{\mathbf{t}}_{dx} = \sum_{k\in s_d} a_k\mathbf{x}_k$. By Taylor linearization, $\hat{t}_{d\mathrm{GREG}}$ is approximated by a population fit regression estimator

$\hat{t}_{d\mathrm{REG}} = \hat{t}_{d\mathrm{HT}} + (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx})'\mathbf{B}_d$ applied in U_d. The estimator $\hat{t}_{d\mathrm{REG}}$ is unbiased for t_d, and so the GREG estimator is nearly unbiased. Although GREG incorporates a model, it is model-assisted, not model-dependent, because the model only yields a fixed population quantity $\mathbf{B}_d$, and GREG is nearly design unbiased even when the model is not valid. By (7), the variance of $\hat{t}_{d\mathrm{GREG}}$ can be estimated using sample residuals $e_k = y_k - \mathbf{x}'_k\hat{\mathbf{B}}_d$:

$$\hat{V}_1(\hat{t}_{d\mathrm{GREG}}) = \sum_{k\in s_d}\sum_{l\in s_d}(a_k a_l - a_{kl})e_k e_l. \tag{9}$$

The GREG estimator can be written as a weighted sum of observations incorporating so-called g-weights:

$$\hat{t}_{d\mathrm{GREG}} = \sum_{k\in s_d} a_k g_{dk} y_k;\ g_{dk} = I_{dk} + I_{dk}(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx})'\hat{\mathbf{M}}_d^{-1}\mathbf{x}_k,$$

where $\hat{\mathbf{M}}_d = \sum_{i\in s_d} a_i\mathbf{x}_i\mathbf{x}'_i$ and $I_{dk} = I\{k \in U_d\}$ is the domain membership indicator. The g-weights are used in a variance estimator

$$\hat{V}_2(\hat{t}_{d\mathrm{GREG}}) = \sum_{k\in s_d}\sum_{l\in s_d}(a_k a_l - a_{kl})g_{dk}e_k g_{dl}e_l \tag{10}$$

(Hidiroglou and Patak, 2004; Särndal et al., 1989 and 1992, p. 235). In practice, $\hat{V}_1$ and $\hat{V}_2$ often yield similar results but $\hat{V}_2$ in (10) is preferable (Fuller, 2002; Särndal et al., 1989).

3.4. Calibration estimators

Calibration is based on information about known totals of auxiliary variables $\mathbf{x}_k$, also called *benchmark variables*, at an aggregate level. In model-free calibration (Särndal, 2007) discussed here, it is not necessary to impose a model on the data. Suppose the population is divided into *calibration groups* U_c $(c = 1, 2, \ldots, C)$ so that every domain U_d is contained within one of the groups and the population totals $\mathbf{t}_{cx} = \sum_{k\in U_c}\mathbf{x}_k$ of auxiliary variables are known. The domain totals $\mathbf{t}_{dx}$ are not required. Direct *calibration estimator* of the domain total t_d is a weighted sum of observations:

$$\hat{t}_{d\mathrm{CAL}} = \sum_{k\in s_d} w_k y_k,$$

where the *calibration weights* w_k have to satisfy the *calibration equation*s

$$\sum_{k\in s_c} w_k\mathbf{x}_k = \sum_{k\in U_c}\mathbf{x}_k = \mathbf{t}_{cx}$$

for every calibration group. It follows immediately that calibration estimator applied to the auxiliary data yields the known totals. We therefore expect that the weighted sum of y over s_d is close to t_d.

There are two main approaches to calibration, one based on a *distance measure* and the other based on *instrument vectors* (Chapter 25). In the distance measure approach, the weights w_k minimize a distance to the design weights a_k, subject to the calibration equations (Deville and Särndal, 1992; Singh and Mohl, 1996). An example of a calibration estimator incorporating an instrument vector $\mathbf{z}_k$ is

$$\hat{t}_{d\mathrm{CAL}} = \sum_{k \in s_d} a_k (1 + \boldsymbol{\lambda}' \mathbf{z}_k) y_k,$$

where $\boldsymbol{\lambda}' = (\mathbf{t}_{cx} - \hat{\mathbf{t}}_{cx})' \left(\sum_{k \in s_c} a_k \mathbf{z}_k \mathbf{x}'_k \right)^{-1}$. It should be noted that the values of instrument z-variables need to be known only for the sample (or need to be estimated); they are not necessarily treated as proper auxiliary information in the same manner as the auxiliary x-variables. For practical purposes, a natural choice is $\mathbf{z}_k = \mathbf{x}_k$; an optimal choice is discussed in Estevao and Särndal (2004).

As in (7), the variance of $\hat{t}_{d\mathrm{CAL}}$ is estimated by

$$\hat{V}(\hat{t}_{d\mathrm{CAL}}) = \sum_{k \in s_c} \sum_{l \in s_c} (a_k a_l - a_{kl})(y_{dk} - \mathbf{x}'_{ck} \hat{\mathbf{B}}_{cd})(y_{dl} - \mathbf{x}'_{cl} \hat{\mathbf{B}}_{cd}),$$

where $\mathbf{x}_{ck} = I\{k \in U_c\} \mathbf{x}_k$ (Estevao and Särndal, 2006), and

$$\hat{\mathbf{B}}_{cd} = \left(\sum_{k \in s_c} a_k \mathbf{z}_k \mathbf{x}'_{ck} \right)^{-1} \left(\sum_{k \in s_c} a_k \mathbf{z}_k y_{dk} \right).$$

When U_c is much larger than U_d, the variance can become large. Therefore, we should attempt to find a calibration group that agrees closely with the domain of interest.

Our GREG estimator of Section 3.3 is actually a special case of calibration, sometimes called linear calibration estimator, as the weights $a_k g_{dk}$ minimize a certain chi-square distance to design weights a_k, subject to domain-level calibration equations $\sum_{k \in s_d} a_k g_{dk} \mathbf{x}_k = \mathbf{t}_{dx}$.

Calibration is contrasted with GREG estimation in Särndal (2007). Särndal and Lundström (2005) discuss calibration in the context of adjustment for unit nonresponse in sample surveys.

3.5. *Computational example with direct estimation under a planned domain structure*

In this section, we demonstrate with real data the direct Horvitz–Thompson, Hájek, and GREG estimation of totals for domains. The data set contains disposable income of households in $D = 12$ regions of Western Finland. The population consists of $N = 431{,}000$ households. In addition to the income data, the record of a household shows the number of household members who had higher education (variable EDUC) and the number of months in total the household members were employed (EMP) during last year. All three variables were determined using administrative registers. For this computational exercise, we had access to population level information on all variables. This gives a possibility to compare sample estimates to the known population values.

We were interested in the yearly total disposable income $t_d = \sum_{k \in U_d} y_k$ in the regions $U_d (d = 1, \ldots, D)$. A sample of 1000 households was drawn from the population by using stratified πPS (without-replacement type probability proportional to size sampling) with household size as the size variable. To demonstrate estimation for planned domains, we interpret here the sample as a stratified sample where the regions constitute the strata. Thus, the domain structure is of planned type, where the regional sample sizes are considered fixed by the sampling design. In Section 4.2, we use the same sample in estimation for unplanned domains, where the regional sample sizes are considered random.

In Table 2, we grouped the domains by sample size into minor ($8 \leq n_d \leq 33$), medium-sized ($34 \leq n_d \leq 45$) and major ($46 \leq n_d \leq 277$) domains, where n_d is the observed domain sample size in domain U_d. There were four domains in each domain size class.

Results are shown in Table 2. The absolute relative error of an estimator in domain d is calculated as $|\hat{t}_d - t_d|/t_d$ and domain group's MARE is the mean of absolute relative errors over domains in the group. Correspondingly, MCV is the mean coefficient of variation of the estimate over domain group. The coefficient of variation is calculated as $\text{s.e}(\hat{t}_d)/\hat{t}_d$, where s.e refers to the estimated standard error of an estimator. For variance estimation, we approximated the design by with-replacement type probability-proportional-to-size sampling (PPS). The variance estimators for ordinary HT (column 1) and the Hájek type estimator (column 2) were defined by (4) and (6), respectively. The Hájek estimator, which contains the known domain sizes N_d, yielded better results than ordinary HT.

A calibration estimator, the direct GREG estimator with linear assisting model,

$$Y_k = \beta_{0d} + \beta_{1d}\text{EMP}_k + \varepsilon_k \text{(column 3) or}$$
$$Y_k = \beta_{0d} + \beta_{1d}\text{EMP}_k + \beta_{2d}\text{EDUC}_k + \varepsilon_k \text{(column 4)},$$

and variance estimator (10) incorporated the known domain sizes and domain totals of EMP (column 3) and EDUC (column 4). The model parameters were estimated by WLS with weights $a_k = 1/\pi_k$. By GREG, we obtained clearly smaller MARE and MCV figures than by HT.

Adding information in the estimation procedure improved the results until the assisting model contained both EMP and EDUC: inclusion of EDUC in GREG decreased MCV but average errors did not always decrease. In large domains, the average error and MCV were usually smaller than in small domains.

Table 2
Mean absolute relative error (MARE) and mean coefficient of variation (MCV) of direct HT, Hájek, and calibration (GREG) estimators of totals for minor, medium-sized, and major domains by using various amounts of auxiliary information in a planned domains case

	HT		Hájek		Calibration (GREG)			
Auxiliary Information	1 None		2 Domain Sizes		3 Domain Sizes and Domain Totals of EMP		4 Domain Sizes and Domain Totals of EMP and EDUC	
Domain sample size class	MARE (%)	MCV (%)	MARE (%)	MCV (%)	MARE (%)	MCV (%)	MARE (%)	MCV (%)
Minor $8 \leq n_d \leq 33$	11.5	11.9	5.3	10.9	5.8	7.7	6.4	6.8
Medium $34 \leq n_d \leq 45$	7.6	9.0	6.4	9.0	3.7	8.0	3.6	8.1
Major $46 \leq n_d \leq 277$	12.5	5.2	4.7	5.6	4.3	4.7	5.2	3.7

4. Indirect estimators in domain estimation

4.1. Generalized regression estimators

4.1.1. Linear GREG

Indirect estimators use y-values also from other domains than the domain of interest. While direct estimators can be derived from corresponding estimators for population, indirect estimators require new results. This holds for unplanned domain structures in particular, but the methodology below applies also to planned domains when indirect estimators are used, for example, when the GREG estimator is assisted by a model fitted to the whole sample. Thus, direct estimators can be treated as a special case of indirect estimators. If the auxiliary information is not available at the domain level but at a higher aggregate level, or if the population frame does not include domain membership data, the calibration approach might be preferred to GREG.

We first assume a common linear fixed-effects regression model $Y_k = \mathbf{x}'_k\boldsymbol{\beta} + \varepsilon_k$ for all domains. The corresponding population fit parameter $\mathbf{B}$ (Section 3.2) is estimated as in Section 3.2. The linear GREG estimator of domain total t_d incorporates fitted values $\hat{y}_k$ of the common model:

$$\hat{t}_{d\text{GREG}} = \sum_{k\in U_d} \hat{y}_k + \sum_{k\in s_d} a_k(y_k - \hat{y}_k), \tag{11}$$

where $\hat{y}_k = \mathbf{x}'_k\hat{\mathbf{B}}$. In general, this is an indirect estimator, since all y-values in the sample contribute.

There is a whole spectrum of model types describing various assumptions about differences between domains (e.g., Lehtonen et al., 2003). If the domains are assumed similar enough, the model may contain only intercept and slopes common to all domains. At the other end of the spectrum, the model is equivalent to a set of separate models for each domain, and all estimators are of direct type. A more parsimonious model might have separate parameters for the largest domains and common parameters for the small domains. It is also possible to use a model formulation with domain-specific intercepts and common slopes or nonlinear model formulations (e.g., Lehtonen et al., 2005). These extensions are discussed in Section 5.

In (11), unit-level auxiliary information about $\mathbf{x}_k$, also including known domain membership, for all population units is assumed. Actually, since the assisting model for (11) is linear, GREG estimation does not require unit-level information on $\mathbf{x}_k$. It is enough to have access to the vector $\mathbf{t}_{dx}$ of domain totals of auxiliary variables in the population and the corresponding HT estimates $\hat{\mathbf{t}}_{dx}$ in the sample. This can be seen by writing the GREG estimator in the standard textbook form,

$$\hat{t}_{d\text{GREG}} = \hat{t}_{d\text{HT}} + (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx})'\hat{\mathbf{B}}.$$

An alternative calibration form incorporates g-weights:

$$\hat{t}_{d\text{GREG}} = \sum_{k\in s} a_k g_{dk} y_k,$$

where $g_{dk} = I_{dk} + (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx})'\hat{\mathbf{M}}^{-1}\mathbf{x}_k$, $I_{dk} = I\{k \in U_d\}$, and $\hat{\mathbf{M}} = \sum_{i\in s} a_i\mathbf{x}_i\mathbf{x}'_i$. The g-weights are often small outside domain sample s_d.

The variance of $\hat{t}_{d\text{GREG}}$ is estimated by a double sum over the whole sample s:

$$\hat{V}(\hat{t}_{d\text{GREG}}) = \sum_{k\in s}\sum_{l\in s}(a_k a_l - a_{kl})g_{dk}e_k g_{dl}e_l \tag{12}$$

(Särndal et al., 1992, p. 401). Alternatively, the sum extends only over the domain sample s_d (Hidiroglou and Patak, 2004). For the direct estimator, these two forms are identical. These variance estimators do not take into account that the sample size n_{s_d} for an unplanned domain is random. To account for the randomness, we might apply GREG assisted by a model fitted to the extended domain variables $y_{dk} = I\{k \in U_d\}y_k$ (Estevao et al., 1995). It has also been proposed to fit the model to the original y_k and replace the residuals e_k in the variance estimator by "extended residuals" $e_{dk} = I\{k \in U_d\}y_k - \hat{y}_k$ (Lehtonen and Pahkinen, 2004, p. 202; Särndal, 2001, p. 39).

The basic direct and indirect GREG estimators and their variance estimators for the case of planned domains, discussed this far, are presented in Table 3 below. In both GREG estimators, access to domain-level auxiliary totals of x-variables is assumed. A key difference is in the model formulation: the direct GREG estimator employs domain-specific assisting models, whereas a model common for all domains is postulated for the indirect GREG estimator. Direct GREG estimation uses domain sample data in variance estimation; the data use extends to the whole sample in indirect GREG.

The GREG estimator (11) has been modified to take into account the domain size N_d when known:

$$\hat{t}_{d\text{GREG}(N)} = \sum_{k\in U_d}\hat{y}_k + (N_d/\hat{N}_d)\sum_{k\in s_d}a_k(y_k - \hat{y}_k) = \sum_{k\in s}a_k g_{dk(N)}y_k, \tag{13}$$

where $g_{dk(N)} = (N_d/\hat{N}_d)I_{dk} + (\mathbf{t}_{dx} - (N_d/\hat{N}_d)\hat{\mathbf{t}}_{dx})'\hat{\mathbf{M}}^{-1}\mathbf{x}_k$ and $\hat{N}_d = \sum_{k\in s_d}a_k$. This estimator has smaller variance than the estimator (11) because the weighted mean of the residuals is more stable. The variance estimator of $\hat{t}_{d\text{GREG}(N)}$ contains the weights $g_{dk(N)}$ instead of g_{dk}. If inference is conditional on observed sample domain sizes, $\hat{t}_{d\text{GREG}(N)}$ is conditionally nearly unbiased, whereas the ordinary GREG is conditionally biased (Hidiroglou and Patak, 2004; Särndal and Hidiroglou, 1989). Therefore, $\hat{t}_{d\text{GREG}(N)}$ yields better conditional confidence intervals. On the other hand, domain estimators (13) are not additive; their sum is not usually equal to the GREG estimator of the population total.

Table 3
The basic direct and indirect GREG estimators and their variance estimators for the planned domains case

	GREG Estimator Type	
	Direct	Indirect
Model formulation	$Y_k = \mathbf{x}'_k\boldsymbol{\beta}_d + \varepsilon_k$	$Y_k = \mathbf{x}'_k\boldsymbol{\beta} + \varepsilon_k$
GREG estimator	$\hat{t}_{d\text{GREG}} = \hat{t}_{d\text{HT}} + (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx})'\hat{\mathbf{B}}_d$	$\hat{t}_{d\text{GREG}} = \hat{t}_{d\text{HT}} + (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx})'\hat{\mathbf{B}}$
Variance estimator	$\hat{V}(\hat{t}_{d\text{GREG}}) = \sum_{k\in s_d}\sum_{l\in s_d}(a_k a_l - a_{kl})g_{dk}e_k g_{dl}e_l$	$\hat{V}(\hat{t}_{d\text{GREG}}) = \sum_{k\in s}\sum_{l\in s}(a_k a_l - a_{kl})g_{dk}e_k g_{dl}e_l$

4.1.2. Composite estimation for domains

As noted in Section 2.2.5, domains with small sample size can present problems in design-based estimation. This also holds for the GREG estimator (11). For example, the GREG estimate for a small domain is not necessarily bounded within an acceptable range. Even when only positive y-values are valid, the GREG estimate may be negative for a small domain when a negative residual is associated with a large weight a_k. In addition, although the GREG estimator (11) is nearly design unbiased, its design variance becomes large for a small domain. *Composite* or *combined* estimators have been proposed to overcome these kinds of problems.

Consider a composite estimator $\hat{t}_{d\text{COMB}} = \lambda_d \hat{t}_{d\text{GREG}} + (1 - \lambda_d)\hat{t}_{d\text{SYN}}$, which is constructed as a weighted sum of the design-based GREG estimator (11) and a model-based synthetic estimator $\hat{t}_{d\text{SYN}} = \sum_{k \in U_d} \hat{y}_k = \sum_{k \in U_d} \mathbf{x}'_k \hat{\mathbf{B}}$. The domain-specific weight $\lambda_d (0 \leq \lambda_d \leq 1)$ is chosen such that λ_d is close to one for large domains and approaches zero with decreasing domain sample size n_{s_d}. Thus, for small domains, the estimator $\hat{t}_{d\text{COMB}}$ will be close to the synthetic estimator $\hat{t}_{d\text{SYN}}$; the GREG estimator (11) will be obtained when the domain sample size is large. Different strategies in choosing λ_d are possible, leading to composite estimators of optimal type or sample size dependent type (see Rao, 2003a, Section 4.3).

The rationale behind composite estimation is obvious. The composite estimator can be written as $\hat{t}_{d\text{COMB}} = \hat{t}_{d\text{SYN}} + \lambda_d \sum_{k \in s_d} a_k (y_k - \hat{y}_k)$, reproducing the GREG estimator (11) with $\lambda_d = 1$. The design variance is of order $O(n^{-1})$ for the synthetic term $\hat{t}_{d\text{SYN}}$ and of order $O\big(n_{s_d}^{-1}\big)$ for the bias correction term $\sum_{k \in s_d} a_k (y_k - \hat{y}_k)$. If the domain sample size n_{s_d} is large, the weight λ_d should be close to one and a sufficiently small variance will be obtained for $\hat{t}_{d\text{COMB}}$. For a small domain, the variance of the correction factor of the GREG will be large and it is beneficial to decrease the value of λ_d because the variance of the component $\hat{t}_{d\text{SYN}}$ tends to be small. This is an example of "trading bias against variance": by suitable choice of λ_d, a balance between the potential design bias of the synthetic estimator and the instability of the GREG estimator is achieved. The price to be paid for the variance reduction is increased design bias because the synthetic estimator $\hat{t}_{d\text{SYN}}$ is generally design biased. The MSE of the composite estimator will be smaller than the MSE of the GREG estimator if the underlying model is not too bad for the given domain. However, with a poor-fitting model, the bias component of the MSE can dominate, leading to increased MSE.

An example of a *sample size dependent composite estimator* is provided by the GREG estimator (13), with $\lambda_d = N_d / \hat{N}_d$. We noted in Section 4.1.1 that the GREG estimate (13) is not necessarily bounded within an acceptable range. The likelihood of this occurrence is reduced when $N_d / \hat{N}_d$ is replaced by $\hat{N}_d / N_d$ in a domain where $n_{s_d} < \sum_{k \in U_d} \pi_k$ (Hidiroglou and Särndal, 1985). Further, Särndal and Hidiroglou (1989) proposed an estimator called *dampened regression estimator* given by

$$\hat{t}_{d\text{DRE}} = \sum_{k \in U_d} \hat{y}_k + (\hat{N}_d / N_d)^{c-1} \sum_{k \in s_d} a_k (y_k - \hat{y}_k),$$

where $c = 0$ if $\hat{N}_d \geq N_d$ and $c = 2$ if $\hat{N}_d < N_d$.

Variants of composite estimators have often been used in practice. Examples of early references are Schaible et al. (1977), and Kumar and Lee (1983). A method called regression composite estimation is discussed in the context of repeated surveys, such as a Labour Force Survey, in Singh et al. (1994), Bell (2001), Fuller and Rao (2001), Gambino et al. (2001), and Singh et al. (2001). Design-based composite estimation, including MSE estimation, is discussed more extensively in Rao (2003a). Model-based composite estimation is treated in Chapter 32.

4.1.3. Model groups approach

Instead of using a common model fitted to the whole sample, it is sometimes more convenient to consider a set of regression models defined for nonoverlapping subsets $U_p(p = 1, 2, \ldots, P)$ of the population called *model groups* (Estevao et al., 1995). In regional classification, there is often a hierarchy of regions, and model groups are larger regions composed of domains. More generally, the boundaries of the sets U_p do not have to agree with domain boundaries, and interleaving is allowed. In model group U_p, we define a model $Y_k = \mathbf{x}'_{kp}\boldsymbol{\beta}_p + \varepsilon_k;\ k \in U_p$. Here, the vectors $\mathbf{x}_{kp}$ may contain different variables in different groups U_p. Naturally, this ensemble of models is equivalent with a single regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$, where $\mathbf{X} = \text{diag}(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_P)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \ldots, \boldsymbol{\beta}'_P)'$. The general theory of the GREG estimator applies, but the $\mathbf{X}$ matrix is perhaps impractically large. It is easier to consider the separate models. For that purpose, the sample is divided into subsets $s_p = U_p \cap s$ and further into sets $s_{pd} = s_p \cap s_d$. If we know the auxiliary totals $\mathbf{t}_{dpx} = \sum_{k \in U_p \cap U_d} \mathbf{x}_{kp}$, estimated by $\hat{\mathbf{t}}_{dpx} = \sum_{k \in s_{pd}} a_k \mathbf{x}_{kp}$, the domain total GREG estimator (11) can be written as

$$\hat{t}_{d\text{GREG}} = \sum_p \sum_{k \in s_{pd}} \hat{y}_k + \sum_p \sum_{k \in s_{pd}} a_k(y_k - \hat{y}_k) = \hat{t}_{d\text{HT}} + \sum_p (\mathbf{t}_{dpx} - \hat{\mathbf{t}}_{dpx})'\hat{\mathbf{B}}_p,$$

where $\hat{\mathbf{B}}_p$ is obtained by fitting the regression model in model group U_p:

$$\hat{\mathbf{B}}_p = \hat{\mathbf{M}}_p^{-1} \sum_{i \in s_p} a_i \mathbf{x}_{ip} y_i \tag{14}$$

and $\hat{\mathbf{M}}_p = \sum_{i \in s_p} a_i \mathbf{x}_{ip} \mathbf{x}'_{ip}$.

The model groups approach can be generalized by the use of *overlapping* sets $U_{p(d)}$ that are defined for each domain U_d so that $U_d \subset U_{p(d)}$. In regional statistics, an example of $U_{p(d)}$ is the neighborhood of a region U_d, the union of U_d and all neighboring regions sharing a common border with the region. This makes sense if the neighboring regions are similar due to spatial correlations (e.g., D'Alo et al., 2006; Petrucci et al., 2005). Since there is no single regression model that is equivalent to the ensemble of separate regression models, the estimators are not necessarily additive.

When the models are defined separately for each domain ($U_p = U_d$), the resulting estimator is direct. In small domains, the direct estimator typically has large variance. Therefore, it has been common to use indirect estimator assisted by a model fitted in a larger subset of the sample. Design-based estimation with an indirect estimator is challenged by Estevao and Särndal (2004) but indirect estimation might be useful at least for the small domains. Hidiroglou and Patak (2004) note that an indirect estimator (13) incorporating $\hat{N}_d$ may be preferred to a corresponding direct estimator when the domain sample size is very small.

The auxiliary totals are not always known in every domain but only in the model groups U_p. This situation can be addressed by calibration. An alternative is the calibration-type GREG estimator discussed in Estevao et al. (1995). It is necessary to fit the regression models to the extended domain variables $y_{dk} = I\{k \in U_d\}y_k$:

$$\hat{t}_{d\mathrm{GREG}(G)} = \sum_{k\in s} a_k y_{dk} + \sum_{p} (\mathbf{t}_{px} - \hat{\mathbf{t}}_{px})'\hat{\mathbf{B}}_{p(d)}, \tag{15}$$

where the auxiliary total over U_p is denoted by $\mathbf{t}_{px}$, its HT estimate by $\hat{\mathbf{t}}_{px}$, and $\hat{\mathbf{B}}_{p(d)} = \hat{\mathbf{M}}_p^{-1} \sum_{i\in s_p} a_i \mathbf{x}_{ip} y_{di}$. Only model groups U_p that intersect the domain are included in (15). An alternative expression for unit-level auxiliary data is

$$\hat{t}_{d\mathrm{GREG}(G)} = \sum_{k\in U} \hat{y}_{dk} + \sum_{k\in s} a_k \big(y_{dk} - \hat{y}_{dk}\big),$$

where $\hat{y}_{dk} = \mathbf{x}'_{kp}\hat{\mathbf{B}}_{p(d)}$ for $k \in U_p$. The calibration equations hold at the model group level, that is, the total estimates of auxiliary variables agree with the known totals over U_p. This approach is adopted, for example, in GES and CLAN software packages.

The variance estimator for (15) is calculated using all residuals $e_{dk} = y_{dk} - \mathbf{x}'_{kp}\hat{\mathbf{B}}_{p(d)}$:

$$\hat{V}(\hat{t}_{d\mathrm{GREG}(G)}) = \sum_{k\in s}\sum_{l\in s} (a_k a_l - a_{kl}) g_{p(k)k} e_{dk} g_{p(l)l} e_{dl}, \tag{16}$$

with $g_{pk} = 1 + (\mathbf{t}_{px} - \hat{\mathbf{t}}_{px})'\hat{\mathbf{M}}_p^{-1}\mathbf{x}_{kp}$ and $k \in U_{p(k)}$ (Estevao et al., 1995; Hidiroglou and Patak, 2004). Obviously, the regression models fitted to y_{dk} will not fit the data well in a large model group and the residuals are often large. This inflates the variance; the problem is often met if a model group contains several domains.

4.1.4. A general class of domain estimators

Estevao and Särndal (2004) define a general class of estimators including both GREG estimators and calibration estimators based on an instrument vector: suppose the auxiliary totals are known over sets U_p, called calibration groups or model groups. For practical purposes, we again assume that the error variance σ_k^2 is constant. The regression parameter is estimated using subpopulations U_m and U_l:

$$\hat{\mathbf{B}}_{ml} = \left(\sum_{k\in s} a_k \mathbf{z}_k I_{mk} \mathbf{x}'_k\right)^{-1} \left(\sum_{k\in s} a_k \mathbf{z}_k I_{lk} y_k\right),$$

where $I_{mk} = I\{k \in U_m\}$ and $I_{lk} = I\{k \in U_l\}$, and $\mathbf{z}_k$ is an instrument vector, in GREG chosen as $\mathbf{z}_k = \mathbf{x}_k$. The domain estimator for $U_d \subset U_p$ is $\hat{t}_d = \hat{t}_{d\mathrm{HT}} + (\mathbf{t}_{px} - \hat{\mathbf{t}}_{px})'\hat{\mathbf{B}}_{ml}$, where the estimators $\hat{t}_{d\mathrm{HT}}$ and $\hat{\mathbf{t}}_{px}$ are HT estimators of the population totals of y_{dk} and $\mathbf{x}_{kp} = I\{k \in U_p\}\mathbf{x}_k$, respectively. As special cases, the calibration estimator based on the instrument vectors $\mathbf{z}_k$ has $U_m = U_p$ and $U_l = U_d$, as well as the GREG estimator incorporating model groups U_p. The ordinary GREG (11) has $U_m = U_l$. In GREG, the regression model is fitted to the whole sample (when $U_l = U$), to each domain (when $U_l = U_d$) or to calibration groups (when $U_l = U_p$). All these estimators are design consistent, and their relative bias tends to zero as $O(n^{-1/2})$.

Estevao and Särndal (2004) show that the design variance of the estimator is minimized by choosing $U_m = U_p$, $U_l = U_d$, and $\mathbf{z}_k = \sum_{l\in U} (a_k a_l/a_{kl} - 1) I_{pl}\mathbf{x}_l$. These instrument variables are estimated by $\mathbf{z}_k = a_k^{-1} \sum_{l\in s} (a_k a_l - a_{kl})\, I_{pl}\mathbf{x}_l$. The resulting estimator is then essentially identical with the so-called *optimal estimator* (Montanari, 1987; Montanari and Ranalli, 2002; Rao, 1994), which minimizes the design variance (Estevao and Särndal, 2004, p. 656)

$$\begin{aligned}\mathrm{Var}(\hat{t}_d) &= \mathrm{Var}(\hat{t}_{d\mathrm{HT}} + (\mathbf{t}_{px} - \hat{\mathbf{t}}_{px})'\mathbf{B})\\ &= \mathrm{Var}(\hat{t}_{d\mathrm{HT}}) + \mathbf{B}'\mathrm{Var}(\hat{\mathbf{t}}_{px})\mathbf{B} - 2\mathbf{B}'\mathrm{Cov}(\hat{t}_{d\mathrm{HT}}, \hat{\mathbf{t}}_{px})\end{aligned}$$

with respect to $\mathbf{B}$. Unfortunately, the optimal estimator is often unstable, especially for designs more complex than SRS (Estevao and Särndal, 2004, p. 657). In practice, we should probably use $\mathbf{z}_k = I_{pk}\mathbf{x}_k$ instead. Then the estimator is the GREG estimator based on model groups. Note that the optimal estimator is a direct estimator using the y-values only from the given domain ($U_l = U_d$). The ordinary GREG estimator has approximately the same asymptotic variance as the optimal calibration estimator only if $U_p = U_d$. Andersson and Thorburn (2005) discuss optimality of a calibration estimator in relation to GREG estimation.

4.1.5. One-stage and two-stage designs

In addition to element-level sampling designs discussed so far, we can define GREG estimators for clusters (Estevao et al., 1995). In single-stage cluster sampling, a sample s_C of clusters is first drawn with design weights a_i^C and all elements in each sample cluster are surveyed. Clusters are grouped into model groups $C_p (p = 1, 2, \ldots, P)$. Consider a cluster $i \in C_p$ with elements s_i and auxiliary data $\mathbf{x}_i$. A regression model is defined for the sum y_{di}^C of y-variables $y_{dk} = I_{dk} y_k$ over the cluster:

$$y_{di}^C = \sum_{k\in s_i} y_{dk} = \mathbf{x}'_i \boldsymbol{\beta}_p + \varepsilon_i,$$

where the error variance is $\mathrm{Var}(\varepsilon_i) = \sigma_i^2$. The regression parameter is estimated for group C_p by

$$\hat{\mathbf{B}}_p = \hat{\mathbf{M}}_p^{-1} \sum_{i\in s_C \cap C_p} a_i^C \mathbf{x}_i y_{di}^C/\sigma_i^2,$$

where $y_{di}^C = \sum_{k\in s_i} y_{dk}$ and $\hat{\mathbf{M}}_p = \sum_{i\in s_C\cap C_p} a_i^C \mathbf{x}_i \mathbf{x}'_i/\sigma_i^2$.

The error variance $\mathrm{Var}(\varepsilon_i)$ can hardly be assumed constant, but, for example, it can often be assumed to be proportional to the size n_i of the cluster: $\sigma_i^2 = n_i\sigma^2$. Then the unknown σ^2 cancels out from $\hat{\mathbf{B}}_p$.

Using known auxiliary totals $\mathbf{t}_{px}^C = \sum_{i\in C_p} \mathbf{x}_i$ and their estimates $\hat{\mathbf{t}}_{px}^C = \sum_{i\in s_C\cap C_p} a_i^C \mathbf{x}_i$, we estimate t_d by

$$\hat{t}_{d\mathrm{GREG}(C)} = \sum_p \sum_{i\in s_c\cap C_p} a_i^C g_{pi}^C y_{di}^C,$$

where $g_{pi}^C = 1 + \left(\mathbf{t}_{px}^C - \hat{\mathbf{t}}_{px}^C\right)' \hat{\mathbf{M}}_p^{-1} \mathbf{x}_i/\sigma_i^2$.

The variance of $\hat{t}_{d\mathrm{GREG}(C)}$ is estimated using residuals $e_{di} = y_{di}^C - \mathbf{x}'_i\hat{\mathbf{B}}_p$ and the inclusion probabilities of clusters:

$$\hat{V}(\hat{t}_{d\mathrm{GREG}(C)}) = \sum_{i \in s_C}\sum_{j \in s_C}(a_i^C a_j^C - a_{ij}^C)g_{p(i)i}^C e_{di} g_{p(j)j}^C e_{dj}$$

with $i \in C_{p(i)}$ and $j \in C_{p(j)}$.

In two-stage sampling, the first-stage sample consists of primary sampling units (PSU), such as clusters. Then in each sample PSU, a sample of elements is drawn. The design weight of element k is a product $a_k = a_i^C a_{k|i}$ of the weight a_i^C of PSU i and the conditional design weight $a_{k|i}$ of element k within PSU i. This generalizes to more stages. If the model groups are defined at the PSU level, the regression models define how the PSU totals depend on auxiliary variables. However, the PSU totals are not known, and we use their HT estimates $\hat{t}_{di} = \sum_{k \in s_i} a_{k|i} y_{dk}$ instead. The GREG estimator of the domain total is

$$\hat{t}_{d\mathrm{GREG}(2)} = \sum_{p}\sum_{i \in s_c \cap C_p} a_i^C g_{pi}^C \hat{t}_{di}$$

but variance estimation requires more complex derivations (e.g., Estevao et al., 1995). Falorsi et al. (2000) discuss some simple estimation methods under two-stage sampling and Estevao and Särndal (2006) discuss calibration under two-stage and two-phase sampling.

4.2. *Computational example with direct and indirect estimation under an unplanned domain structure*

Domain totals are estimated here by direct Horvitz–Thompson and indirect GREG estimators. We use the same sample as in Section 3.5. This allows a comparison of results with the case of direct estimation for planned domains. There were $D = 12$ regions (domains) in our population. To demonstrate domain estimation for unplanned domains, we recognize that the regional sample sizes n_{s_d} are not fixed in the sampling design but are random (in Section 3.5, we assumed a case of planned domains with domain sample sizes fixed by stratification).

In addition to the income data for households, the sample data set includes the variables EDUC (number of household members who had higher education) and EMP (the number of months in total the household members were employed during last year). We again estimate the domain totals of disposable income of households in the 12 regions. We use the same auxiliary data as in Section 3.5. In addition to direct HT, we computed two indirect GREG estimates. Results are shown in Table 4. MARE is the mean absolute relative error and MCV is the mean coefficient of variation of the estimate over domain group.

The variance of ordinary HT (column 1 in Table 4) was estimated by $\hat{V}_U(\hat{t}_{d\mathrm{HT}})$ (5). As expected, in the present case of unplanned domains, the HT estimator had larger MCV than in the case of planned domains (column 1 in Table 2). The random domain sample size increased the variance of domain estimators.

In GREG, we first illustrate the model groups approach. We assumed that the population size N and the population total of EMP only were known. We thus had a single model

Table 4
Mean absolute relative error (MARE) and mean coefficient of variation (MCV) of HT and indirect GREG estimators of totals for minor, medium-sized, and major domains by using various amounts of auxiliary information in an unplanned domains case

	HT		GREG			
Auxiliary Information	1 None		2 Population Size and EMP Total		3 Domain Sizes and Domain Totals of EMP	
Domain sample size class	MARE (%)	MCV (%)	MARE (%)	MCV (%)	MARE (%)	MCV (%)
Minor $8 \le n_{s_d} \le 33$	11.5	28.3	11.5	28.3	7.6	9.0
Medium $34 \le n_{s_d} \le 45$	7.6	20.3	7.4	20.3	3.8	8.1
Major $46 \le n_{s_d} \le 277$	12.5	9.6	12.5	9.4	4.1	5.0

group, that is, the whole population. The indirect GREG estimator (15) was assisted by model

$$Y_{dk} = \beta_0 + \beta_1 \mathrm{EMP}_k + \varepsilon_k.$$

We thus did not use domain-level auxiliary information. For each domain, we fitted the model to the extended domain variables $y_{dk} = I\{k \in U_d\} y_k$. The variables y_{dk} were also included in the variance estimator (16). This GREG estimator (column 2) did not yield smaller errors or MCV than the HT estimator. The population level information was not powerful for domain estimation in this case, confirming the argument of favoring the use of lower level aggregates of auxiliary variables if available (Estevao and Särndal, 2004).

The second indirect GREG estimator (column 3) was assisted by a common model

$$Y_k = \beta_0 + \beta_1 \mathrm{EMP}_k + \varepsilon_k$$

fitted to the whole sample, and domain sizes and domain totals of EMP were assumed known. The variance was estimated using (12). This estimator outperformed the other three estimators. The MCV was larger than in the comparable direct GREG estimator for planned domains (column 3 in Table 2), as expected. The use of extended domain residuals $e_{dk} = y_{dk} - \hat{y}_k$ in the variance estimator would have affected the MCV only slightly. Increasing the number of auxiliary variables in GREG did not yield further improvement. The size correction with known domain size (13) resulted in small decrease in average errors, but MCV increased slightly.

We did have access to several cross-sectional yearly data sets of the survey and the corresponding auxiliary data. With two last year's data, the domains were defined by cross classification of year and region, yielding altogether 24 domains. We fitted models containing the year and interactions of year with EMP and EDUC, but the results did not improve. A model fitting the whole sample better does not necessarily fit better for the data in domains of interest, and even if it did, a better fitting model does not guarantee better GREG estimates in one particular sample although improvement is expected on an average.

The problem of model choice is discussed in Lehtonen and Veijanen (1998), Estevao and Särndal (1999), Hedlin et al. (2001), Lehtonen et al. (2003, 2005), and Hidiroglou and Patak (2004). We address model choice in GREG estimation further in Sections 5.1 and 5.2.

4.3. *Ratios and percentiles for domains*

4.3.1. *Ratios and means*

Consider estimating the *ratio* $R_d = t_{dy}/t_{dz}$ of two unknown totals $t_{dy} = \sum_{k \in U_d} y_k$ and $t_{dz} = \sum_{k \in U_d} z_k$. An example is the unemployment rate, which is the ratio of the number of unemployed and the size of the labor force in the domain. Another example is the proportional area of fields allocated to, say, wheat in a region, estimated using data obtained from each farm k; we only need to know (y_k, z_k) for units in the sample from area d. A simple, nearly unbiased estimator of R_d is $\hat{R}_d = \hat{t}_{dy}/\hat{t}_{dz}$. We denote the ratio of two HT estimators by $\hat{R}_{d\mathrm{HT}}$ and the ratio of two GREG estimators by $\hat{R}_{d\mathrm{GREG}}$.

In a case of planned domains, the variance estimators for the ratios of direct HT and GREG estimators are defined as follows (Särndal et al., 1992, p. 178, 296):

$$\hat{V}(\hat{R}_{d\mathrm{HT}}) = \frac{1}{\hat{t}^2_{dz\mathrm{HT}}} \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl})(y_k - \hat{R}_{d\mathrm{HT}} z_k)(y_l - \hat{R}_{d\mathrm{HT}} z_l),$$

$$\begin{aligned}\hat{V}(\hat{R}_{d\mathrm{GREG}}) = \frac{1}{\hat{t}^2_{dz\mathrm{GREG}}} &\sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl}) g_{dk} (e_{yk} - \hat{R}_{d\mathrm{GREG}} e_{zk}) \\ &\times g_{dl}(e_{yl} - \hat{R}_{d\mathrm{GREG}} e_{zl}),\end{aligned}$$

where the residuals $e_{yk} = y_k - \mathbf{x}'_k \hat{\mathbf{B}}_{dy}$ and $e_{zk} = z_k - \mathbf{x}'_k \hat{\mathbf{B}}_{dz}$ are obtained from regression models fitted in the domain to y_k and z_k, respectively, and the g-weights are common to both models. In the case of indirect GREG,

$$\begin{aligned}\hat{V}(\hat{R}_{d\mathrm{GREG}}) = \frac{1}{\hat{t}^2_{dz\mathrm{GREG}}} &\sum_{k \in s} \sum_{l \in s} (a_k a_l - a_{kl}) g_{dk} (e_{yk} - \hat{R}_{d\mathrm{GREG}} e_{zk}) \\ &\times g_{dl}(e_{yl} - \hat{R}_{d\mathrm{GREG}} e_{zl}),\end{aligned}$$

where $e_{yk} = y_k - \mathbf{x}'_k \hat{\mathbf{B}}_y$ and $e_{zk} = z_k - \mathbf{x}'_k \hat{\mathbf{B}}_z$ are residuals of models fitted in the whole sample.

With unplanned domains, we can estimate the domain ratio by the ratio of two population level estimators using extended domain variables $y_{dk} = I\{k \in U_d\} y_k$ and $z_{dk} = I\{k \in U_d\} z_k$. In the case of HT, this ratio is actually identical with $\hat{R}_{d\mathrm{HT}}$ defined above:

$$\hat{R}_{d(e)} = \frac{\sum_{k \in s} a_k y_{dk}}{\sum_{k \in s} a_k z_{dk}}.$$

Moreover, $\hat{V}(\hat{R}_{d(e)}) = \hat{V}(\hat{R}_{d\mathrm{HT}})$. In contrast, the variance estimator of a ratio of two GREG estimators incorporating the extended domain variables is

$$\hat{V}(\hat{R}_{d\mathrm{GREG}}) = \frac{1}{\hat{t}^2_{dz\mathrm{GREG}}} \sum_{k\in s} \sum_{l\in s} (a_k a_l - a_{kl}) g_{dk}(e_{ydk} - \hat{R}_{d\mathrm{GREG}} e_{zdk}) \\ \times g_{dl}(e_{ydl} - \hat{R}_{d\mathrm{GREG}} e_{zdl}),$$

where $e_{ydk} = y_{dk} - \hat{y}_{dk}$ and $e_{zdk} = z_{dk} - \hat{z}_{dk}$ are from models fitted to the extended domain variables.

Domain mean $\bar{y}_d = t_d/N_d$ can be estimated by $\hat{\bar{y}}_d = \hat{t}_d/N_d$ when the domain size N_d is known. The variance estimator is correspondingly $\hat{V}(\hat{t}_d)/N_d^2$. An alternative is to interpret the domain mean as a ratio $R_d = t_d/t_{dz}$, where $t_{dz} = \sum_{k\in U_d} z_k = N_d$ is defined for $z_k = I_{dk} = I\{k \in U_d\}$. The estimator $\hat{t}_{dz}$ is an estimator of the domain size: $\hat{t}_{dz} = \hat{N}_d = \sum_{k\in s_d} a_k$. This is applicable also when N_d is unknown. The mean estimator is then $\hat{R}_d = \hat{t}_d/\hat{t}_{dz}$, and the variance is estimated by the formula for $\hat{V}(\hat{R}_d)$ with $z_k = I_{dk}$. Comparison of estimators of domain means is studied in Särndal et al. (1992), p. 412.

The *ratio estimator* is an estimator of t_d based on $\hat{R}_d$ and a known total t_{dz}: $\hat{t}_{dR} = t_{dz}\hat{R}_d$. It is nearly unbiased for t_d and its variance is estimated by $t_{dz}^2 \hat{V}(\hat{R}_d)$. If the domain size N_d is known, a ratio estimator of t_d derived from an estimator of the domain mean is $\hat{t}_{d(N)} = N_d \hat{\bar{y}}_d$. If $\hat{\bar{y}}_d$ is estimated by $\hat{\bar{y}}_d = \hat{t}_{d\mathrm{HT}}/\hat{N}_d$, then $\hat{t}_{d(N)}$ is a special case of the Hájek type estimator. The estimates $\hat{t}_{d(N)}$ do not, in general, add up to the estimate of the population total.

4.3.2. *Percentile estimation for domains*

Percentiles, such as median and quartiles, are important in certain surveys, notably surveys of income statistics including median household income, income deciles, and derived poverty measures. The percentiles can be estimated using an estimated distribution function (Chambers and Dunstan, 1986; Chambers and Tzavidis, 2006; Rao et al., 1990; Tzavidis et al., 2007); recently, calibration has been used (Rueda et al., 2007a; Särndal, 2007; Wu and Sitter, 2001a). Harms and Duschene (2006) use known percentiles of auxiliary variables. These studies have not considered estimation of domain percentiles, but most population estimators can be apparently generalized for domain estimation. We also suggest straightforward application of the estimation equation approach of Binder and Patak (1994). Percentile estimation is discussed also in Chapter 36.

The distribution function is defined for a finite population domain U_d of size N_d as

$$F_d(t) = \sum_{k\in U_d} I(y_k \le t)/N_d,$$

where the indicator function $I(y_k \le t)$ equals 1 when $y_k \le t$ and 0 otherwise. The pth percentile is $\theta = \theta(p) = \inf\{t : F_d(t) \ge p\}$, that is, we find the smallest value θ for which proportion p of the y_k are smaller than or equal to θ. In the finite population, we choose percentiles among the values y_k. Then the percentile is $\theta = \min\{y_k : F_d(y_k) > p; k \in U_d\}$. It is useful to restate the problem as follows: the solution of $F_d(\theta) = p$ satisfies an estimating equation defined for $u(y, \theta) = I(y \le \theta) - p$:

$W_d(\theta) = \int_{-\infty}^{\infty} u(y, \theta) dF_d(y) = 0$. As F_d is a step function, the equation is

$$W_d(\theta) = \sum_{k \in U_d} u(y_k, \theta)/N_d = \sum_{k \in U_d} (I(y_k \leq \theta) - p)/N_d = 0.$$

When using a sample without auxiliary information, we estimate $W_d(\theta)$ by HT and use equation

$$\hat{W}_d(\theta) = \sum_{k \in s_d} a_k (I(y_k \leq \theta) - p)/N_d = 0.$$

This has the same form as the optimal estimating function of Godambe and Thompson (1986a), although their theory seems to require differentiable $u(y, \theta)$. The solution satisfies

$$\hat{F}_{d\mathrm{HT}}(\theta) = \sum_{k \in s_d} a_k I\,(y_k \leq \theta)/\hat{N}_d = p.$$

The function $\hat{F}_{d\mathrm{HT}}(\theta)$ is interpreted as an HT estimator of the distribution function. It is monotone, nondecreasing, and bounded in [0, 1]. This simplifies finding the percentile. The smallest value y_k for which $\hat{F}_{d\mathrm{HT}}(y) > p$ is found from sorted data in the same way as in the binary search algorithm. Percentile searching can be more complicated if the estimated distribution function is not monotone. Rao et al. (1990) have suggested that a monotone distribution function estimate is derived by tracking maxima. Rueda et al. (2007a) have presented a calibration-based monotone and nondecreasing estimator of the distribution function.

Särndal et al. (1992, p. 203) give an approximate variance estimator of $\hat{F}_{d\mathrm{HT}}(\theta)$:

$$\hat{V}_{\hat{F}}(\theta) = \hat{V}(\hat{F}_{d\mathrm{HT}}(\theta)) = \frac{1}{\hat{N}_d^2} \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl})(I(y_k \leq \hat{\theta}) - p)(I(y_l \leq \hat{\theta}) - p).$$

Under the assumption of normality of $\hat{F}_{d\mathrm{HT}}(\theta)$ close to p, a 95% confidence interval for $F_d(\theta)$ is $[p - 1.96\hat{V}_{\hat{F}}(\theta)^{1/2}, p + 1.96\hat{V}_{\hat{F}}(\theta)^{1/2}]$. A confidence interval for θ is obtained from the equivalent equality $P\{\hat{F}_{d\mathrm{HT}}^{-1}(p - 1.96\hat{V}_{\hat{F}}(\theta)) \leq \theta \leq \hat{F}_{d\mathrm{HT}}^{-1}(p + 1.96\hat{V}_{\hat{F}}(\theta))\} = 0.95$.

When auxiliary data are available, Binder and Patak (1994) have proposed a generalization of the estimating equation containing $\alpha(\mathbf{x}, \boldsymbol{\beta}, \theta) = E\,(u(y, \theta)|\mathbf{x})$ under a model with parameter $\boldsymbol{\beta}$:

$$\int_{-\infty}^{\infty} \alpha\,(\mathbf{x}, \boldsymbol{\beta}, \theta)\, d[F_{X;d}(\mathbf{x}) - \hat{F}_{X;d}(\mathbf{x})] + \int_{-\infty}^{\infty} u(y, \theta) d\hat{F}_d(y) = 0,$$

where $F_{X;d}$ is the distribution function of $\mathbf{x}$ in domain d. In the case of percentile estimation, $\alpha(\mathbf{x}, \boldsymbol{\beta}, \theta) = P\{Y \leq \theta|\mathbf{x}; \boldsymbol{\beta}\} - p$. Let us denote the probability $P\{Y \leq \theta|\mathbf{x} = \mathbf{x}_k; \boldsymbol{\beta}\}$ by p_k. The estimating equation is

$$\sum_{k \in U_d} \frac{1}{N_d}\,(p_k - p) - \sum_{k \in s_d} \frac{1}{\hat{N}_d} a_k\,(p_k - p) + \sum_{k \in s_d} \frac{1}{\hat{N}_d} a_k\,(I(y_k \leq \theta) - p) = 0.$$

When we substitute estimated probabilities $\hat{p}_k$ for p_k (see below), we obtain an equation

$$\hat{F}_{d\text{GREG}}(\theta) = p \quad \text{with} \quad \hat{F}_{d\text{GREG}}(\theta) = \frac{1}{N_d}\sum_{k\in U_d}\hat{p}_k + \frac{1}{\hat{N}_d}\sum_{k\in s_d} a_k(I(y_k \leq \theta) - \hat{p}_k).$$

This is interpreted as a GREG estimator (13) of the distribution function; the indicators $I(y_k \leq \theta)$ are the observations and $\hat{p}_k$ are the fitted values. This estimator is similar to a difference estimator defined in Rao et al. (1990). It is indirect if the probabilities $\hat{p}_k$ are estimated from the whole sample, and then the percentile should be searched using all observations of the sample, but variance is still probably large in a small domain. We can estimate the variance of $\hat{F}_{d\text{GREG}}(\theta)$ using the ordinary variance estimator $\hat{V}$ of GREG (13). This would yield a confidence interval with end points $\hat{F}^{-1}_{d\text{GREG}}(p - 1.96\hat{V}^{1/2})$ and $\hat{F}^{-1}_{d\text{GREG}}(p + 1.96\hat{V}^{1/2})$, but its properties are not known yet.

The estimates $\hat{p}_k$ are obtained from a logistic regression model fitted to the indicators $I(y_k \leq \theta)$ in the sample, preferably by maximizing a pseudolikelihood that contains design weights. Alternatively, one can obtain $\hat{p}_k$ using the empirical distribution function $\hat{F}_{\hat{e}}$ of the standardized residuals $\hat{e}_k = (y_k - \mathbf{x}'_k\hat{\boldsymbol{\beta}})/\hat{\sigma}$ in the sample; $\hat{p}_k = \hat{F}_{\hat{e}}((\theta - \mathbf{x}'_k\hat{\boldsymbol{\beta}})/\hat{\sigma})$, or by using the fitted values $\hat{y}_k$ in the population; $\hat{p}_k = I\big(\hat{y}_k \leq \theta\big)$ (Rao et al., 1990; Wu and Sitter, 2001a). In domain estimation, it is an open question whether to use only the data in the domain or a larger data set to obtain possibly better estimates $\hat{p}_k$.

5. Extended GREG family for domain estimation

5.1. Assisting models

A fixed-effects linear model is often chosen as an assisting model for a GREG estimator of direct or indirect type; this was the case in Sections 3 and 4. When the model does not fit well in a domain, the population fit residuals $E_k = y_k - \hat{y}_k$ in that domain can be large, inflating the estimator's variance. Nonlinear models may fit better, especially if the variable of interest is binary or multinomial. Mixed models can offer an interesting alternative for direct and indirect GREG estimators with fixed-effects type assisting models. By introducing suitable random components in the model, flexible accounting for the domain differences is allowed. The extended GREG family of domain estimators refers to GREG type estimators where the assisting model is a member of the family of generalized linear mixed models (GLMM; e.g., Breslow and Clayton, 1993; McCulloch and Searle, 2001). Lehtonen and Veijanen (1998) and Lehtonen et al. (2003, 2005) have introduced GREG estimators of the form (11) assisted by logistic, multinomial logistic, and mixed models. This approach might be attractive at least from a modeller's point of view. Torabi and Rao (2008) compare the MSE behavior of an EBLUP estimator with a GREG estimator assisted by a mixed model, introduced in Lehtonen and Veijanen (1999).

Access to reliable auxiliary information is essential for accurate domain estimation. In Sections 3 and 4, we worked with aggregate-level auxiliary data. Now, we assume access at unit-level auxiliary data. Let us assume that the auxiliary vector value $\mathbf{x}_k = (1, x_{1k}, \ldots, x_{jk}, \ldots, x_{Jk})'$ and domain membership is known and specified in the frame for every unit $k \in U$. Consider first a generalized linear fixed-effects model,

$E_m(Y_k) = f(\mathbf{x}_k; \boldsymbol{\beta})$ for a given function $f(\cdot\,; \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ requires estimation, and E_m refers to the expectation under the model. Examples of $f(\cdot\,; \boldsymbol{\beta})$ are a linear functional form or a logistic function. The model fit to the sample data $\{(y_k, \mathbf{x}_k); k \in s\}$ yields the estimate $\hat{\mathbf{B}}$ of $\mathbf{B}$, a finite population counterpart of $\boldsymbol{\beta}$. Using the estimated parameter values, the vector value $\mathbf{x}_k$, and the domain membership of k, we compute the predicted value $\hat{y}_k = f(\mathbf{x}_k; \hat{\mathbf{B}})$ for every $k \in U$, which is possible under our assumptions.

A similar reasoning applies for a generalized linear mixed model involving random effects in addition to the fixed effects. The model specification is $E_m(Y_k|\mathbf{u}_d) = f(\mathbf{x}'_k(\boldsymbol{\beta}+\mathbf{u}_d))$, where $\mathbf{u}_d$ is a vector of random effects defined at the domain level. Using the estimated parameters, predicted values $\hat{y}_k = f(\mathbf{x}'_k(\hat{\mathbf{B}}+\hat{\mathbf{u}}_d))$ are computed for all $k \in U$.

An example of a mixed model formulation is a multinomial logistic mixed model for a binary or polytomous y-variable. In addition to domains U_d, a second subdivision of U arises: for an m-class polytomous variable, the population is also subdivided into classes denoted U_i, $i = 1, \ldots, m$. For class U_i, denote the response variable as y_i with value $y_{ik} = 1$ if $k \in U_i$ and $y_{ik} = 0$ otherwise. We want to estimate the class frequencies or totals $t_{id} = \sum_{k\in U_d} y_{ik}$, $i = 1, \ldots, m$, for all domains U_d. For a binary y-variable ($m = 2$), the domain totals are $t_d = \sum_{k\in U_d} y_k$. The multinomial logistic mixed model is of the form

$$E_m(y_{ik}|\mathbf{u}_d) = P\{y_{ik} = 1|\mathbf{u}_d\} = \frac{\exp(\mathbf{x}'_k(\boldsymbol{\beta}_i + \mathbf{u}_{id}))}{1 + \sum_{r=2}^{m} \exp(\mathbf{x}'_k(\boldsymbol{\beta}_r + \mathbf{u}_{rd}))}$$

for $k \in U_d, i = 1, \ldots, m, d = 1, \ldots, D$, where $\mathbf{x}_k$ is a known vector value for every $k \in U$, $\boldsymbol{\beta}_i$ is a vector of fixed effects common for all domains, $\mathbf{u}_d = (\mathbf{u}'_{1d}, \ldots, \mathbf{u}'_{id}, \ldots, \mathbf{u}'_{md})'$, and $\mathbf{u}_{id}$ is a vector of domain-specific random effects, defined for the classes of the y-variable. To avoid identifiability problems, we set $\boldsymbol{\beta}_1 = 0$. Lehtonen et al. (2005) give special cases of the model.

Obviously, the possible nonlinearity of the model complicates the method. For example, we cannot express the sum of fitted values using the sum of auxiliary variables; in general, $\sum_{k\in U_d} \hat{y}_k \neq \left(\sum_{k\in U_d} \mathbf{x}_k\right)' \hat{\mathbf{B}}$. As a consequence, the GREG estimator cannot be written using the totals of auxiliary variables. The representation incorporating g-weights is also invalid, and the variance estimator with g-weights is not appropriate. For a given model specification, the GREG estimator of domain total $t_d = \sum_{k\in U_d} y_k$ remains the one given by (11), that is, the form $\hat{t}_{d\mathrm{GREG}} = \sum_{k\in U_d} \hat{y}_k + \sum_{k\in s_d} a_k(y_k - \hat{y}_k)$, $d = 1, \ldots, D$. The latter component in GREG, an HT estimator of the residual total, aims at correcting for the bias of the synthetic part.

We could use a simpler variance estimator (9), but it is probably negatively biased. A resampling-based variance estimator might be preferred. Stukel et al. (1996) discuss jackknife type variance estimation for calibration estimators.

For simplicity, we concentrate now on linear models (Lehtonen et al., 2003). The model specification of a linear mixed model is $E_m(Y_k|\mathbf{u}_d) = \mathbf{x}'_k(\boldsymbol{\beta} + \mathbf{u}_d) = (\beta_0 + u_{0d}) + (\beta_1 + u_{1d})x_{1k} + \cdots + (\beta_J + u_{Jd})x_{Jk}$, where $\mathbf{u}_d = (u_{0d}, u_{1d}, \ldots, u_{Jd})'$ is a vector of random effects defined at the domain level. The random effects are assumed to have common distribution. In estimation, they are often shrunken towards zero. The random components of $\mathbf{u}_d$ represent deviations from the corresponding coefficients of the fixed-effects part of the model. In practice, not all components are treated as random; for some j, $u_{jd} = 0$ for every d. A simple example is a model that includes domain-specific random intercepts u_{0d} as the only random term. If all components of $\mathbf{u}_d$ are set to zero, a fixed-effects model is attained. The mixed model is usually estimated by using ML and restricted or residual maximum likelihood (REML) methods (e.g., Goldstein, 2002; McCulloch and Searle, 2001). Using the estimated parameters, predicted values $\hat{y}_k = \mathbf{x}'_k(\hat{\mathbf{B}} + \hat{\mathbf{u}}_d)$ are computed for all $k \in U$. The predictions $\{\hat{y}_k;\, k \in U\}$ differ from one model specification to another.

An additional possible direction for extension of the GREG concept is explored in Breidt and Opsomer (2000). These authors use nonparametric regression techniques to obtain the fitted values necessary for a GREG type estimator. Zheng and Little (2003, 2004) use penalized spline nonparametric mixed models for a similar purpose. Nonparametric and semiparametric estimation is discussed in Chapter 27. By using suitable mixed models, Jiang and Lahiri (2006) introduce a model-assisted empirical best prediction approach for domain means.

5.2. *Computational example for extended GREG family estimators*

We compare empirically the design bias and accuracy of model-assisted GREG type estimators of domain totals of a continuous y-variable for different linear assisting models (fixed-effects, mixed). Results are based on Monte Carlo simulation experiments, where repeated systematic probability proportional-to-size samples (πPS design) were drawn from an artificially generated fixed and finite population. The inclusion probabilities were $\pi_k = nx_{1k} \big/ \sum_{k \in U} x_{1k}$. The weights $a_k = 1/\pi_k$ varied between 54.5 and 599.8. We used unit-level auxiliary data.

In the Monte Carlo experiment, for an estimate $\hat{t}_d(s_v)$ obtained for sample $s_v;\ v = 1, 2, \ldots, K$, we computed for each domain U_d the absolute relative bias (ARB; defined as the ratio of the absolute value of bias to the true value), given by $\left|(1/K)\sum_{v=1}^{K} \hat{t}_d(s_v) - t_d\right|/t_d$, and relative root mean squared error (RRMSE), defined as the ratio of the root MSE to the true value, given by $\sqrt{(1/K)\sum_{v=1}^{K} (\hat{t}_d(s_v) - t_d)^2} \big/ t_d$.

There were $D = 100$ domains in the population. The size of domain U_d was proportional to $\exp(q_d)$, where q_d was simulated from U(0,2.9). We had 47 domains with minor sample sizes, 19 domains with medium sample sizes, and 34 domains with major sample sizes. These three size classes were defined on the basis of expected sample size $n(t_{dx_1}/t_{x_1})$ in domain U_d, where x_1 is the size variable used in πPS sampling. The domain size classes were less than 70, 70–119, and 120 or more units. The smallest domain of the generated population had 1721 units and the largest had 28,614.

The auxiliary variable x_1 was simulated from uniform distribution U(1,11). Another auxiliary variable x_2, unrelated to the sampling design, was simulated from $U(-5, 5)$. The random effects u_d and random slopes ν_{id}, $i = 1, 2$, were simulated for each domain from multinormal distribution with variances $\mathrm{Var}(u_d) = 1$, $\mathrm{Var}(\nu_{id}) = 0.125$ and correlations $\mathrm{Corr}(u_d, \nu_{id}) = -0.5$; $\mathrm{Corr}(\nu_{1d}, \nu_{2d}) = 0$. The error term ε was generated from N(0,100). Values of the y-variable were simulated as $y_k = 1 + (1 + \nu_{1d})x_{1k} + (1 + \nu_{2d})x_{2k} + u_d + \varepsilon_k$. Correlations of the variables in the population were as follows: $\mathrm{corr}(y, x_1) = 0.44$, $\mathrm{corr}(y, x_2) = 0.45$, and $\mathrm{corr}(x_1, x_2) \approx 0$. Domain means of the y-variable were approximately equal but the totals differed considerably: The means of domain totals were 50,977 for minor domains, 131,776 for medium domains, and 263,979 for major domains.

Our population size was $N = 1{,}000{,}000$ and sample size $n = 10{,}000. K = 1000$ independent samples were selected. The following assisting models (groups A, B, C, and D) were considered:

Model A1, $Y_k = \beta_{0d} + \varepsilon_k$, $k \in U_d$, producing a direct estimator GREG-A1.

Model A2, $Y_k = \beta_0 + u_d + \varepsilon_k$, $k \in U$, producing an indirect estimator MGREG-A2.

Model B1, $Y_k = \beta_{0d} + \beta_2 x_{2k} + \varepsilon_k$, $k \in U$, producing an indirect estimator GREG-B1.

Model B2, $Y_k = \beta_0 + u_d + \beta_2 x_{2k} + \varepsilon_k$, $k \in U$, producing an indirect estimator MGREG-B2.

Model C1, $Y_k = \beta_{0d} + \beta_1 x_{1k} + \varepsilon_k$, $k \in U$, producing an indirect estimator GREG-C1.

Model C2, $Y_k = \beta_0 + u_d + \beta_1 x_{1k} + \varepsilon_k$, $k \in U$, producing an indirect estimator MGREG-C2.

Model D1, $Y_k = \beta_{0d} + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$, $k \in U$, producing an indirect estimator GREG-D1.

Model D2, $Y_k = \beta_0 + u_d + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$, $k \in U$, producing an indirect estimator MGREG-D2.

A-models did not contain auxiliary information. In B-models, the auxiliary variable x_2 was used, whereas the πPS size variable x_1 was included in C-models. Both auxiliary variables were included in D-models. Note that for the models A1, A2, B1, and B2, the sampling is *informative* (see Chapter 39), because the values of the y-variable depend on x_1 but the predictor is not included in the model. In models A1, B1, C1, and D1, the domain differences were accounted for by domain-specific fixed effects β_{0d}, and in A2, B2, C2, and D2 by domain-specific random intercepts $\beta_0 + u_d$. We incorporated the design weights a_k in the estimation procedures of model parameters, including the mixed models. This facilitates the condition of "internal bias calibration" (a proper combination of model formulation and estimation procedure under a given sampling design) proposed, for example, by Firth and Bennett (1998). The design weights were included in a REML method introduced in Saei and Chambers (2004) by modifying matrix products of $\mathbf{X}$, $\mathbf{y}$, the $\mathbf{Z}$ matrix whose columns are domain indicators, and $\mathbf{e}$, the vector of residuals: for example, the sample-based $\mathbf{X}'_s\mathbf{X}_s$ in the original algorithm was replaced by $\mathbf{X}'_s\mathbf{W}\mathbf{X}_s$, where $\mathbf{W}$ is the diagonal matrix of design weights. $\mathbf{X}'_s\mathbf{W}\mathbf{X}_s$ is an estimate of the corresponding product $\mathbf{X}'_U\mathbf{X}_U$ defined in the population.

The design bias of GREG estimators remained negligible for all model formulations considered (Table 5). In model groups A, B, C, and D, a mixed model formulation yielded slightly better results than fixed model formulation. Accuracy improved when incorporating in B-type assisting models the auxiliary variable x_2 (which was unrelated to the sampling design). GREG-C1 and GREG-C2 outperformed the A-type and B-type estimators. Best accuracy was obtained for the D-models. Thus, the inclusion of the πPS size variable x_1 in C-type and D-type assisting models appears powerful in this case. This strategy facilitates "double use" (Särndal, 1996) of the auxiliary information (i.e., to use it both in the sampling design and in the estimation phase).

Table 5
Average absolute relative bias (ARB) and average relative root mean squared error (RRMSE) of GREG estimators of domain totals for minor, medium-sized, and major domains of the generated population

Model and Estimator	Average ARB (%)			Average RRMSE (%)		
	Domain Size Class			Domain Size Class		
	Minor (20 – 69)	Medium (70 – 119)	Major (120+)	Minor (20 – 69)	Medium (70 – 119)	Major (120+)
Model A1 $Y_k = \beta_{0d} + \varepsilon_k$						
GREG-A1	1.2	0.7	0.3	20.2	11.9	8.5
Model A2 $Y_k = \beta_0 + u_d + \varepsilon_k$						
MGREG-A2	0.5	0.5	0.3	19.9	11.8	8.5
Model B1 $Y_k = \beta_{0d} + \beta_2 x_{2k} + \varepsilon_k$						
GREG-B1	1.2	0.6	0.3	18.3	10.7	7.7
Model B2 $Y_k = \beta_0 + u_d + \beta_2 x_{2k} + \varepsilon_k$						
MGREG-B2	0.5	0.4	0.2	18.0	10.6	7.7
Model C1 $Y_k = \beta_{0d} + \beta_1 x_{1k} + \varepsilon_k$						
GREG-C1	0.4	0.3	0.2	17.5	10.3	7.5
Model C2 $Y_k = \beta_0 + u_d + \beta_1 x_{1k} + \varepsilon_k$						
MGREG-C2	0.3	0.3	0.2	17.3	10.2	7.5
Model D1 $Y_k = \beta_{0d} + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$						
GREG-D1	0.4	0.3	0.2	15.3	8.8	6.5
Model D2 $Y_k = \beta_0 + u_d + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$						
MGREG-D2	0.3	0.3	0.2	15.1	8.7	6.5

5.3. *Other extensions*

A class of extended generalized regression estimators (EGRE) has been introduced by Montanari and Ranalli (2002) but it has not been applied in domain estimation yet. Calibration has been generalized in various ways. Wu and Sitter (2001a) discuss model calibration approach by defining the calibration equations for the fitted values: $\sum_{k \in s} w_{ks} \hat{y}_k = \sum_{k \in U} \hat{y}_k$. This approach works well with nonlinear models but auxiliary information is needed at unit level. Nonparametric model calibration by neural networks is studied in Montanari and Ranalli (2005), who assumed access to unit-level auxiliary information. Lehtonen et al. (2008) compared model calibration and GREG in the context of domain estimation.

6. Software

6.1. *SAS applications and macros*

SAS procedure SURVEYMEANS can be used in HT estimation for domains (STRATA and DOMAIN statements) under unequal probability sampling. SAS procedure SURVEYFREQ is available for domain analysis of frequency tables. With some additional programming, SAS procedure SURVEYREG yields GREG estimates for domains. Extended domain variables y_d with $y_{dk} = I_{dk} y_k = I\{k \in U_d\} y_k$ can be used for unplanned domain structures. Variance estimation is based on Taylor linearization.

CALMAR (CALibration on MARgins) and CALMAR 2 are calibration-oriented SAS macro programs of INSEE (Caron and Sautory, 2004; Le Guennec and Sautory, 2003). Methods of Deville and Särndal (1992) and Deville et al. (1993), for example, are implemented.

CLAN is a freely available SAS macro developed at Statistics Sweden (Andersson and Nordberg, 1998). CLAN contains GREG and different calibration methods. Variance estimation is based on Taylor linearization.

GES, Statistics Canada's Generalized Estimation System is a domain estimation package including GREG and calibration estimation (Estevao et al., 1995). The same g-weights can be applied to different y-variables and different domains. Variance is estimated by Taylor linearization or jackknife.

Computer software for sample surveys is discussed further in Chapter 13.

6.2. Application Domest

Domest is an interactive Java application developed for the estimation of totals or means for domains and small areas. It uses methods described in Lehtonen et al. (2003) and Saei and Chambers (2004). Domest provides both model-based and design-based domain estimators. Mixed models are incorporated into EBLUP, synthetic estimator, and pseudo EBLUP (Rao, 2003a). Design-based methods include HT and most GREG methods presented in this chapter. GREG estimation is assisted by fixed-effects regression models or mixed models, fitted with or without design weights. Currently, GREG variance estimation allows SRSWOR, Poisson sampling, and πPS with approximated second-order inclusion probabilities (Berger, 2004, 2005b; Hájek, 1964).

A linear regression model is fitted by OLS or WLS, and a mixed model is fitted by ML or REML (Saei and Chambers, 2004). When the fitting of a mixed model incorporates design weights in the same way as in pseudolikelihood estimation, the design bias of EBLUP seems to decrease.

The mixed model can include both area and time effects. The area effects are then assumed independent and time effects have AR(1) correlations. In a mixed model with spatially correlated random effects, the correlation of the random effects associated with regions a and b distance d_{ab} apart is $\mathrm{Corr}(u_a, u_b) \propto \exp(-d_{ab})$. Spatial correlations may improve the predictive power of a synthetic domain estimator. In a domain missing from the sample, the correlation structure yields a nonzero estimate of the associated random effect.

SAS data or text files can be imported into Domest and output tables are saved as text files or added incrementally to an HTML file.

Domest is developed at Statistics Finland by Ari Veijanen with Risto Lehtonen. It is freely available from the authors.

Acknowledgments

Discussions with Carl-Erik Särndal over past years have inspired us in writing the chapter. He also contributed to Section 5.2, as did Mikko Myrskylä of University of Pennsylvania. Comments of two referees and the Editor were of great help in finalizing the materials. We are thankful to Kari Djerf of Statistics Finland for providing us with the real-life survey data for our empirical examples in Sections 3 and 4. Thanks are also due to Statistics Finland and University of Helsinki for kindly supporting this work.