PROGRAM *DOMEST* FOR ESTIMATION FOR DOMAINS AND SMALL AREAS

PART 1: Technical documentation

Dr. Ari Veijanen
Statistics Finland

Prof. Risto Lehtonen
University of Helsinki

Version 1.0

August 2010

Contents

Summary

Domest is a stand-alone interactive Java application developed for the estimation of totals and means for population subgroups or domains and small areas. The program covers selected methods described in Lehtonen, Särndal and Veijanen (2003), Lehtonen and Veijanen (2009) and Saei and Chambers (2004). Domest provides both design-based and model-based domain estimators with the accompanying variance and MSE estimators. Design-based methods include HT and GREG methods presented in Särndal, Swensson and Wretman (1992) and Lehtonen and Veijanen (2009). GREG estimation is assisted by linear fixed effects regression models or linear mixed models, fitted with or without design weights. Currently, GREG variance estimation allows SRSWOR, Poisson sampling, and πPS with approximated second-order inclusion probabilities (Hájek, 1964; Berger, 2004, 2005). Linear mixed models are incorporated into model-based EBLUP, synthetic estimator and pseudo-EBLUP (Rao, 2003). MSE estimation uses methods described in Rao (2003) and Saei and Chambers (2004).

A linear regression model is fitted by OLS or WLS, and a linear 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 tends to decrease.

The mixed model can include area and/or 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 is $Corr(u_a, u_b) \propto \exp(-d_{ab})$ for regions a and b distance d_{ab} apart. Spatial correlations may improve the predictive power of a synthetic domain estimator. In a domain missing from the sample, the correlation structure yields a non-zero estimate of the associated random effect.

Domest has been tested for MS Windows XP, Vista and Windows 7 platforms. Domest needs Java JRE (version 6 or higher). SAS data or text files can be imported into Domest. Output tables are saved as text files or added incrementally to an HTML file. SAS is only needed if the input data set is in SAS format.

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

Contact: ari.veijanen@pp1.inet.fi

Introduction to Domest

1. Preliminaries

1.1. Sampling designs, models and estimators

1.1.1. Sampling designs and data

Domest handles element-level equal and unequal probability sampling designs. Stratification is allowed. Alternatives of sampling designs are:

1. Simple random sampling without replacement (SRSWOR)
2. Stratified SRSWOR
3. πPS (PPSWOR) with fixed overall sample size n
4. Poisson sampling with expected overall sample size n.

The sampling design is defined on page 'Data'. This page also allows the definition of the stratum variable, if appropriate.

The sample data set and the population data set are defined on page 'Data'. The data sets can be imported into Domest in text or SAS format. The sample data set must contain the weight variable (design weights), as to be determined on page 'Data'. Additional requirements are given in Annex 2. There, requirements are given for response variable y, design weights, domain variable, auxiliary x-variables, strata, variables determining subsets for a random effect, time variable and coordinate variables.

Population data set contains information about the auxiliary variables in the population. These data are either at the unit-level or aggregated over the domain. In unit-level data set each observation represents one population unit. Domest also accepts aggregated population data set that contains known totals of auxiliary x-variables over subsets of the population, e.g. domain totals of the x-variables.

Transformations to variables in sample and population data sets can be carried out on page 'Transformations'. For example, domains are created from the original domain (stratum) variable by function 'Classify!' and corresponding indicator variables are created by function 'Indicators'.

1.1.2. Models and estimators

Domest covers the following design-based and model-based estimator types of domain totals:

1. Design-based estimators
    - Horvitz-Thompson (HT) and Hájek estimators
    - Generalized regression (GREG) estimators with assisting linear fixed-effects model or linear mixed model

2. Model-based estimators
    - EBLUP, pseudo EBLUP and weighted EBLUP (EBLUPW) estimators with underlying linear mixed model
    - Synthetic (SYN) estimator with underlying linear fixed-effects model.

The models in GREG, EBLUP and SYN include linear fixed effects and linear mixed models with random intercepts. For mixed models, random time effects or time-varying effects can be included. Spatial mixed models can be postulated if geocoded (coordinate) data are available.

Design weights are usually incorporated by default into a design-based model fitting procedure such as the estimation of a linear fixed-effects model by weighted least squares (WLS) in GREG. This is not necessarily so for model-based estimators. In Domest, we offer the use of design weights as an option for certain model-based estimation procedures. These are WLS in SYN and weighted ML (ML-W) and REML (REML-W) in EBLUP and pseudo EBLUP estimation, see details in Table 1 and Sections 3 and 4).

Linear fixed-effects models can be specified and fitted in an interactive fashion on page 'Fixed-effects model'. Ordinary (unweighted) LS or WLS can be chosen.

Mixed models are created and fitted on page ‘Linear mixed model’. There, random intercepts, random area effects and/or random time effects can be defined. Estimation is carried out optionally by ML or REML or by the weighted versions ML-W or REML-W.

Domain estimation is carried out on page ‘Domain estimation’. There, a number of operations and options can be chosen:

1. Definition of domains
   - Options for domains: Estimation for domains or estimation for whole population
   - Options for domain type: Unplanned domains or planned domains
   -
2. Model selection for GREG, EBLUP and SYN
   - Options for model choice: Linear regression model or linear mixed model
   - Options for weights: To use or not to use weights in model fitting
   -
3. Selection of statistics
   - Options for statistics: Domain totals and/or domain means
   - Options for the calculation of domain means
   - Options for accuracy measures: Variance (for design-based estimators) and MSE with variance components (for model-based EBLUP estimators)
   - Estimated random effects can be obtained.

A summary of different choices is presented in Table 1.

Technical derivations are presented in Sections 2 and 3. An illustrative example is in Part 2 of the software documentation.

Table 1. Summary of estimators for domains: Estimator and model types and the use of weights in the estimation procedure.

Domain estimator	Weights in domain estimator	Assisting / Underlying model	Weights in model fitting	Model fitting method
Design-based estimators				
HT	Yes	None	-	-
Hájek	Yes	None	-	-
GREG	Yes	Linear fixed-effects model	Yes	WLS
GREG-OLS	Yes	Linear fixed-effects model	No	OLS
MGREGW	Yes	Linear mixed model	Yes	ML-W or REML-W
MGREG	Yes	Linear mixed model	No	ML or REML
Model-based estimators				
SYN	No	Linear fixed-effects model	No	OLS
SYNW	No	Linear fixed-effects model	Yes	WLS
EBLUP(Y)	No	Linear mixed model	No	ML of REML
EBLUP(Y)W	No	Linear mixed model	Yes	ML-W or REML-W
EBLUP(μ)	No	Linear mixed model	No	ML or REML
EBLUP(μ)W	No	Linear mixed model	Yes	ML-W or REML-W
Pseudo-EBLUP	Yes	Linear mixed model	No	ML or REML
Pseudo-EBLUPW	Yes	Linear mixed model	Yes	ML-W or REML-W

2. Design-based estimators of domain totals

2.1. Notation

Consider a fixed and finite population U with N elements indexed by k ($k = 1,...,N$). Population subgroups or domains are indexed by d ($d = 1,...,D$). In population, domains of interest are denoted as $U_d \subset U$. Two different domain structures are allowed. For *unplanned domains*, a single sample s is drawn from U and the corresponding domain samples are $s_d = U_d \cap s \subset U$. For *planned domains* referring to stratified sampling, a separate sample $s_d \subset U_d$ is drawn from each domain (stratum) and $s = \bigcup_{d=1}^{D} s_d$. The population domain sizes are N_d and the domain sample sizes are n_d, where $\sum_{d=1}^{D} N_d = N$ and $\sum_{d=1}^{D} n_d = n$. For planned domains, n_d are fixed by the allocation scheme used in stratification, and N_d are assumed known. For an unplanned domains structure, the domain sample sizes n_d are random, and N_d are assumed known or unknown; in the latter case they are estimated as the sum of design weights over the domain.

The responses are denoted by y_k. The so-called *extended domain variables of interest* or extended domain responses y_{dk} are defined as $y_{dk} = y_k$ for $k \in U_d$ and $y_{dk} = 0$ for $k \notin U_d$. In other words, $y_{dk} = I\{k \in U_d\} y_k$. Because the unknown domain totals are $t_d = \sum_{k \in U_d} y_k = \sum_{k \in U} y_{dk}$, we can estimate the domain total of y in domain d by estimating the population total of y_d. The known population vector values of auxiliary x-variables are denoted by $\mathbf{x}_k = (x_{1k}, x_{2k}, ..., x_{pk})'$. In estimation we usually insert a constant value 1 as the first element of vector $\mathbf{x}_k$.

The inclusion probabilities are $\pi_k = P\{k \in s\}$, and their inverses are design weights $a_k = 1/\pi_k$. The second-order inclusion probabilities $\pi_{ij} = P\{i \in s, j \in s\}$ determine weights $a_{ij} = 1/\pi_{ij}$. For Poisson sampling, $\pi_{ij} = \pi_i \pi_j$, $i \neq j$. In general, $\pi_{ij} = \pi_i$ for $i = j$.

2.2. Horvitz-Thompson (HT) estimator

HT does not require population data or a model, although in Domest it is necessary to specify a simple model from which the y-variable is obtained. Domain total is estimated as the weighted sum of *y*-values over the domain:

$$\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k \ , \tag{1}$$

where the weights are inverses of inclusion probabilities ($a_k = 1/\pi_k$). An HT estimator of domain mean is $\hat{t}_{dHT} / N_d$ or $\hat{t}_{dHT} / \hat{N}_d$, where $\hat{N}_d = \sum_{k \in s_d} a_k$.

Variance estimator of (1) depends on the type of domains. When domains are of *planned* type (domains are the strata in the sampling design and the domain sample sizes n_d are considered fixed), the variance of HT estimator of domain total is estimated by

$$\hat{V}(\hat{t}_{dHT}) = \sum_{i,j \in s_d} \left(a_i a_j - a_{ij} \right) y_i y_j \ . \tag{2}$$

The variance estimator (2) is impractical because it contains second-order inclusion probabilities π_{kl} whose computation is often laborious for practical purposes. Hájek (1964) and Berger (2004, 2005) proposed approximations to π_{kl} of type $\pi_{kl} \approx \pi_k \pi_l [1-(1-\pi_k)(1-\pi_l) m_d^{-1}]$ 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

$$\hat{V}\left(\hat{t}_{dHT}\right) = \sum_{k \in s_d} c_k e_k^2 \ , \tag{3}$$

where $c_k = (1-\pi_k)$ and $e_k = a_k y_k - (\sum_{i \in s_d} c_i)^{-1} \sum_{i \in s_d} c_i a_i y_i$.

Hansen-Hurwitz type variance estimator is available as an option when the domains are of planned type. It is based on approximating the actual design by a with-replacement (WR) type design:

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

When domains are *unplanned* (the domain sample sizes n_d are not fixed in the sampling design but are random), the variance estimator is

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

where y_d are the extended domain variables with values y_{dk} and n is the size of the whole sample data set.

2.3. Hájek estimator

Hájek estimator requires known domain sizes N_d in the population. Known N_d are typically assumed for planned type domain structures where the strata constitute the domains of interest. The domain total is calculated from the weighted domain mean of responses:

$$\hat{t}_{dHajek}=N_d\hat{\bar{y}}_d;\qquad \hat{\bar{y}}_d=\frac{\sum_{k\in s_d}a_k y_k}{\sum_{k\in s_d}a_k} \tag{6}$$

Because the weighted mean is more stable than the weighted sum of responses, Hajek estimator may have smaller variance than HT. For planned domains, the variance estimator is

$$\hat{V}\left(\hat{t}_{dHajek}\right)=\left(\frac{N_d}{\sum_{i\in s_d}w_i}\right)^2\sum_{k\in s_d}\sum_{l\in s_d}\left(a_k a_l-a_{kl}\right)\left(y_k-\hat{\bar{y}}_d\right)\left(y_l-\hat{\bar{y}}_d\right). \tag{7}$$

2.4. Generalized Regression Estimator GREG

2.4.1. GREG assisted by a fixed-effects model

GREG estimator has been traditionally assisted by a domain-specific linear regression model $Y_k = \mathbf{x}_k'\boldsymbol{\beta}_d + \varepsilon_k$ with $Var(\varepsilon_k) = \sigma_k^2$. The model should be fitted using the design weights $a_k = 1/\pi_k$. Assuming constant error variance σ_k^2, the domain-specific population parameter $\mathbf{B}_d$ defined for U_d is estimated by

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

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

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k) = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k e_k . \tag{8}$$

Variance of GREG is estimated by

$$\hat{V}(\hat{t}_{dGREG}) = \sum_{i,j \in s} \left(a_i a_j - a_{ij} \right) g_{di} e_i g_{dj} e_j , \tag{9}$$

where g_{di} are called g-weights. When the model has been fitted separately in each domain (as is often the case for planned domains), the GREG estimator is called *direct* and the g-weights in (9) depend on auxiliary information as follows:

$$g_{dk} = I_{dk} + I_{dk} \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx} \right)' \hat{\mathbf{M}}_d^{-1} \mathbf{x}_k ,$$

where $I_{dk} = I\{k \in U_d\}$ is the domain membership indicator, $\mathbf{t}_{dx} = \sum_{k \in U_d} \mathbf{x}_k$ is the vector of population totals of *x*-variables, $\hat{\mathbf{t}}_{dx} = \sum_{k \in s_d} a_k \mathbf{x}_k$ is the corresponding vector of estimated totals, and

$$\hat{\mathbf{M}}_d = \sum\nolimits_{i \in s_d} a_i \mathbf{x}_i \mathbf{x}_i' .$$

More generally, the model is fitted in the whole sample yielding an *indirect* GREG estimator. Indirect estimators are often used for unplanned domains. The model is of the form

$$Y_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k = \beta_0 + \beta_1 x_{1k} + \ldots + \beta_p x_{pk} + \varepsilon_k$$

and the g-weights in (9) are

$$g_{dk} = I_{dk} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\mathbf{M}}^{-1}\mathbf{x}_k$$

with $\hat{\mathbf{M}} = \sum_{i \in s} a_i \mathbf{x}_i \mathbf{x}_i'$. The g-weights are often small outside domain sample s_d. This form applies also to the direct GREG estimator, if the domain structure is of unplanned type.

Domain differences can be described by including domain indicators in the fixed effects model. All indicators cannot be included as they sum up to vector of ones, which is always in the model, so you have to remove one of the indicators.

For the selection of an assisting model to Domest, it is advisable to carry out a more complete exploratory regression analysis by suitable software such as SAS procedure REG or SURVEYREG.

2.4.2. MGREG estimator assisted by mixed model

The GREG estimator assisted by a linear mixed model is called mixed-model assisted GREG, abbreviated MGREG. The assisting model is of the form

$$Y_k = \mathbf{x}_k'\boldsymbol{\beta} + u_d + \varepsilon_k = \beta_0 + u_d + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon_k . \qquad (10)$$

The domain-specific random intercepts u_d allow variation around the fixed-type intercept β_0. Equations (8) and (9) apply with predictions $\hat{y}_k$ obtained from the mixed model; estimated random effects contribute to the predictions. By default, Domest calculates the variances of MGREG as in (9). Additional technical details on mixed models are included in Section 3.

2.4.3. Options for estimation

Domest page 'Domain estimation' provides the following options for GREG, MGREG and SYN estimation, to be used (instead of the defaults (8) and (9)) under unplanned domain structures:

- Option 1: Use Known Domain Sizes (in a domain size correction)
- Option 2: Use Extended Domain Responses
- Option 3: Use Extended Residuals
- Option 4: Variance Over Domain Only (estimating variance by a double sum over domain).

Option 1

When the *domain size correction* is used, the GREG estimator is

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \frac{N_d}{\sum_{k \in s_d} w_k} \sum_{k \in s_d} a_k \left(y_k - \hat{y}_k \right). \tag{11}$$

The g-weights are now

$$g_{dk} = g_{dk(N)} = I_{dk} c_d + \left(\mathbf{t}_{dx} - c_d \hat{\mathbf{t}}_{dx} \right)' \hat{\mathbf{M}}^{-1} \mathbf{x}_k ,$$

where $c_d = \dfrac{N_d}{\sum_{k \in s_d} a_k}$ and $\hat{\mathbf{M}} = \sum_{i \in s} a_i \mathbf{x}_i \mathbf{x}_i'$ is defined at the whole sample level.

This form is useful when the domains are of unplanned type, but it is not used for planned domains. As the weighted average of residuals is more stable than the sum, estimator (11) may have smaller variance than the ordinary GREG (8). Variance estimator is now

$$\hat{V}\left(\hat{t}_{dGREG} \right) = \sum_{k \in s} \sum_{l \in s} (a_k a_l - a_{kl}) g_{dk(N)} e_k g_{dl(N)} e_l ,$$

where $e_k = y_k - \hat{y}_k$.

A drawback of the domain size correction is that domain estimates are not additive: their sum does not equal the GREG estimate for the whole population. In contrast, the ordinary GREG estimator (8) without the domain size correction is additive over domains.

Option 2

Extended domain responses are $y_{dk} = I\{k \in U_d\} y_k$. If this option is selected, each domain has its own regression model fitted to extended domain responses y_{dk} instead of the original y-variable values y_k. Extended domain responses should be taken into account in the choices of other options: the domain size correction should not used (because N_d are unknown) and the variance is preferably estimated as a sum over the whole sample:

$$\hat{V}\left(\hat{t}_{dGREG}\right) = \sum_{k \in s} \sum_{l \in s} (a_k a_l - a_{kl}) g_{dk} e_{dk} g_{dl} e_{dl} ,$$

where residuals are $e_{dk} = y_{dk} - \hat{y}_{dk}$ and $y_{dk} = I_{dk} y_k$.

In the case of unplanned domains, the resulting GREG variance estimator may take random domain sizes into account better than the other variance estimators.

Option 3

Extended residuals are defined as differences of extended responses and predictions. They are used only in variance estimation:

$$\hat{V}\left(\hat{t}_{dGREG}\right) = \sum_{k \in s} \sum_{l \in s} (a_k a_l - a_{kl}) g_{dk} e^*_{dk} g_{dl} e^*_{dl} ,$$

where $e^*_{dk} = y_{dk} - \hat{y}_k$ are the extended residuals.

Option 4

If the model has been fitted to the whole sample, the variance of the resulting indirect GREG estimator is often calculated over the whole sample (this is the default for unplanned domains). Domest has option of estimating the *variance only over each domain*:

$$\hat{V}(t_{d;GREG}) = \sum_{i,j \in s_d} \left(a_i a_j - a_{ij} \right) g_{di} e_i g_{dj} e_j \ . \tag{12}$$

This usually gives smaller variance estimates than the sum over the whole sample.

The options can be specified for an estimator on Domest page 'Domain estimation'. It should be noted that not all options 1 to 4 can be chosen for an estimator of domain total, as shown in Table 2.

Table 2. Options available in estimation procedure for the various estimators of domain totals under unplanned domain structures.

Estimator	Option			
	1	2	3	4
GREG	Yes	Yes	Yes	Yes
GREG-OLS	Yes	Yes	Yes	Yes
MGREG	Yes	No	Yes	Yes
MGREGW	Yes	No	Yes	Yes
SYN	No	Yes	No	No
SYNW	No	Yes	No	No

Option 1: Domain size correction
Option 2: Extended domain responses
Option 3: Extended residuals
Option 4: Variance only over each domain

A summary of GREG estimators of domain totals and their variance estimators is presented in Table 3, covering direct and indirect estimators for both planned and unplanned domain structures.

Table 3. Direct and indirect GREG estimators and default and optional design-based variance estimators for planned and unplanned domain structures.

GREG estimator type	
Direct	**Indirect**
Model formulation	
Linear fixed-effects model $Y_k = \mathbf{x}_k'\boldsymbol{\beta}_d + \varepsilon_k$	Linear fixed-effects model $Y_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k$ (1) Linear mixed model (random intercepts) $Y_k = \mathbf{x}_k'\boldsymbol{\beta} + u_d + \varepsilon_k$ (2)
GREG estimators	
$\hat{t}_{dGREG} = \sum_{k\in U_d} \hat{y}_k + \sum_{k\in s_d} a_k (y_k - \hat{y}_k)$ $\hat{y}_k = \mathbf{x}_k'\hat{\mathbf{B}}_d$ $\hat{t}_{dGREG} = \sum_{k\in s_d} a_k g_{dk} y_k$ $g_{dk} = I_{dk} + I_{dk}\left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\mathbf{M}}_d^{-1}\mathbf{x}_k$ $\hat{t}_{dGREG(N)} = \sum_{k\in s_d} a_k g_{dk(N)} y_k$ $g_{dk(N)} = \left(N_d / \hat{N}_d\right) I_{dk} + I_{dk}\left(\mathbf{t}_{dx} - \left(N_d / \hat{N}_d\right)\hat{\mathbf{t}}_{dx}\right)' \hat{\mathbf{M}}_d^{-1}\mathbf{x}_k$ $\hat{N}_d = \sum_{k\in s_d} a_k,\ \hat{\mathbf{M}}_d = \sum_{i\in s_d} a_i \mathbf{x}_i \mathbf{x}_i',\ I_{dk} = I\{k \in U_d\}$	$\hat{t}_{dGREG} = \sum_{k\in U_d} \hat{y}_k + \sum_{k\in s_d} a_k (y_k - \hat{y}_k)$ $\hat{y}_k = \mathbf{x}_k'\hat{\mathbf{B}}$ for (1), $\hat{y}_k = \mathbf{x}_k'\hat{\mathbf{B}} + \hat{u}_d$ for (2) $\hat{t}_{dGREG} = \sum_{k\in s} a_k g_{dk} y_k$ $g_{dk} = I_{dk} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\mathbf{M}}^{-1}\mathbf{x}_k$ $\hat{t}_{dGREG(N)} = \sum_{k\in s} a_k g_{dk(N)} y_k$ $g_{dk(N)} = \left(N_d / \hat{N}_d\right) I_{dk} + \left(\mathbf{t}_{dx} - \left(N_d / \hat{N}_d\right)\hat{\mathbf{t}}_{dx}\right)' \hat{\mathbf{M}}^{-1}\mathbf{x}_k$ $\hat{N}_d = \sum_{k\in s_d} a_k,\ \hat{\mathbf{M}} = \sum_{i\in s} a_i \mathbf{x}_i \mathbf{x}_i',\ I_{dk} = I\{k \in U_d\}$
Default variance estimators	
Planned domains $\hat{V}\left(\hat{t}_{dGREG}\right) = \sum_{k\in s_d}\sum_{l\in s_d} (a_k a_l - a_{kl}) g_{dk} e_k g_{dl} e_l$ $e_k = y_k - \hat{y}_k$	Planned domains $\hat{V}\left(\hat{t}_{dGREG}\right) = \sum_{k\in s}\sum_{l\in s} (a_k a_l - a_{kl}) g_{dk} e_k g_{dl} e_l$ $e_k = y_k - \hat{y}_k$
Unplanned domains $\hat{V}\left(\hat{t}_{dGREG}\right) = \sum_{k\in s}\sum_{l\in s} (a_k a_l - a_{kl}) g_{dk} e_k g_{dl} e_l$ $e_k = y_k - \hat{y}_k$	Unplanned domains $\hat{V}\left(\hat{t}_{dGREG}\right) = \sum_{k\in s}\sum_{l\in s} (a_k a_l - a_{kl}) g_{dk} e_k g_{dl} e_l$ $e_k = y_k - \hat{y}_k$
Optional variance estimators for unplanned domains	
Extended responses and predictions $\hat{V}\left(\hat{t}_{dGREG}\right) = \sum_{k\in s}\sum_{l\in s} (a_k a_l - a_{kl}) g_{dk} e_{dk} g_{dl} e_{dl}$ $e_{dk} = y_{dk} - \hat{y}_{dk},\ y_{dk} = I_{dk} y_k$	Extended responses and predictions $\hat{V}\left(\hat{t}_{dGREG}\right) = \sum_{k\in s}\sum_{l\in s} (a_k a_l - a_{kl}) g_{dk} e_{dk} g_{dl} e_{dl}$ $e_{dk} = y_{dk} - \hat{y}_{dk},\ y_{dk} = I_{dk} y_k$
Extended residuals $\hat{V}\left(\hat{t}_{dGREG}\right) = \sum_{k\in s}\sum_{l\in s} (a_k a_l - a_{kl}) g_{dk} e_{dk}^* g_{dl} e_{dl}^*$ $e_{dk}^* = y_{dk} - \hat{y}_k$	Extended residuals $\hat{V}\left(\hat{t}_{dGREG}\right) = \sum_{k\in s}\sum_{l\in s} (a_k a_l - a_{kl}) g_{dk} e_{dk}^* g_{dl} e_{dl}^*$ $e_{dk}^* = y_{dk} - \hat{y}_k$
Known domain sizes $\hat{V}\left(\hat{t}_{dGREG}\right) = \sum_{k\in s}\sum_{l\in s} (a_k a_l - a_{kl}) g_{dk(N)} e_k g_{dl(N)} e_l$ $e_k = y_k - \hat{y}_k$	Known domain sizes $\hat{V}\left(\hat{t}_{dGREG}\right) = \sum_{k\in s}\sum_{l\in s} (a_k a_l - a_{kl}) g_{dk(N)} e_k g_{dl(N)} e_l$ $e_k = y_k - \hat{y}_k$

3. Model-based estimators of domain totals

3.1. EBLUP estimators

3.1.1. Definition

EBLUP, empirical best linear unbiased predictor, is based on a linear mixed model. Recall that under the mixed model,

$$Y_k = \mathbf{x}_k'\boldsymbol{\beta} + \mathbf{z}_k'\mathbf{u} + \varepsilon_k \tag{13}$$

where $\mathbf{u}$ is the vector of random effects (intercepts and slopes) and $\mathbf{z}_k$ is a vector of zeroes and ones chosen so that $\mathbf{z}_k'\hat{\mathbf{u}}$ consists of the correct elements of $\mathbf{u}$ for observation *k*. If there are only area (domain) effects, then $\mathbf{z}_k'\hat{\mathbf{u}} = \hat{u}_d$ is the estimated area effect for domain *d* containing the element *k*. If there are both regional and time effects, $\mathbf{z}_k'\hat{\mathbf{u}} = \hat{u}_d + \hat{v}_{t(k)}$ contains the time effect $\hat{v}_{t(k)}$ for the time period $t(k)$.

Rao (2003, p. 96) considers the estimation of conditional expectations of domain sums of the study variable given the random effects:

$$\mu_d = E\left(\sum_{k\in U_d} Y_k \mid \mathbf{u}\right) = \sum_{k\in U_d} E(Y_k \mid \mathbf{u}) = \sum_{k\in U_d} \left(\mathbf{x}_k'\boldsymbol{\beta} + \mathbf{z}_k'\mathbf{u}_d\right).$$

This is a random variable as it depends on the random effects which are random variables. A model unbiased estimator $\hat{\mu}_d$ of μ_d has expectation

$$E\left(\hat{\mu}_d\right) = E\left(\mu_d\right).$$

The MSE of $\hat{\mu}_d$ is

$$MSE\left(\hat{\mu}_d\right) = E\left(\hat{\mu}_d - \mu_d\right)^2.$$

We denote the EBLUP estimator of μ_d by EBLUP(μ):

$$\hat{t}_{dEBLUP(\mu)} = \hat{\mu}_d = \sum_{k \in U_d} \left(\mathbf{x}_k' \hat{\boldsymbol{\beta}} + \mathbf{z}_k' \hat{\mathbf{u}}_d \right). \quad (14)$$

The estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ in (14) are derived by first considering BLUP estimation of μ_d with known variance components σ_u^2 and σ^2. The final EBLUP estimator is obtained by substituting estimated variance components for their true values.

If we are more interested in estimating the actual domain total

$$t_d = \sum_{k \in U_d} Y_k = \sum_{k \in U_d - s_d} Y_k + \sum_{k \in s_d} Y_k \ ,$$

we might prefer a prediction estimator containing the observed sample values (Saei and Chambers, 2004). The unobserved part of t_d in $U_d - s_d$ remains to be estimated as in EBLUP(μ). We denote the estimator by EBLUP(Y):

$$\hat{t}_{dEBLUP(Y)} = \sum_{k \in U_d - s_d} (\mathbf{x}_k' \hat{\boldsymbol{\beta}} + \mathbf{z}_k' \hat{\mathbf{u}}_d) + \sum_{k \in s_d} y_k \ . \quad (15)$$

The MSE of (15) is defined as

$$MSE\left(\hat{t}_d\right) = E\left(\hat{t}_d - t_d\right)^2.$$

Lehtonen, Myrskylä, Särndal and Veijanen (2007) have introduced design weights into the estimation of model (13). Design weights can be used in the fitting of the mixed model (see page 'Mixed model' in Domest). See details in Section 3.1.2.

The MSE estimator of EBLUP (Rao, 2003; Saei and Chambers, 2004) consists of four terms which are included in Domest output: $MSE = g_1 + g_2 + 2g_3 + g_4$. In the MSE of EBLUP(Y) some population domain sums in the MSE of EBLUP(μ) are replaced by sums over $U_d - s_d$, because the sample domain sum of observations cancels out from the MSE. Therefore EBLUP(Y) tends to have smaller MSE estimates than EBLUP(μ).

3.1.2. Fitting the mixed model

The mixed model has been fitted by iterative ML or REML algorithms (Saei and Chambers, 2004). Lehtonen et al. (2007) introduced design weights into the estimation by equating the error variance to the design variance. The modified algorithms are called ML-W and REML-W. Saei and Chambers (2004) give the necessary theory for the case when the covariance matrix of errors differs from $Cov(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$.

A tentative theoretical argument for incorporating design weights into the estimation is as follows: Consider a linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon};\ Cov(\mathbf{u}) = \boldsymbol{\Omega};\ Cov(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$$

The sample vector can be interpreted as a vector

$$\left(y_1^*, y_2^*, ..., y_N^*\right); y_k^* = I_k y_k = I_k \mathbf{x}_k' \boldsymbol{\beta} + I_k \mathbf{z}_k' \mathbf{u} + I_k \varepsilon_k,$$

where the sample membership indicators I_k are assumed independent of the effects $\mathbf{u}$ and the errors ε_k. The errors $\varepsilon_k^* = I_k \varepsilon_k$ have covariance matrix

$$C = Cov(I_1\varepsilon_1, I_2\varepsilon_2, ..., I_N\varepsilon_N)$$

with elements $[C]_{ii} = \sigma^2 \pi_i$ and $[C]_{ij} = \sigma^2 \pi_{ij}$. It is approximated by a diagonal matrix

$$W = \sigma^2 diag(\pi_1, \pi_2, ..., \pi_N).$$

The covariance of $I_i \mathbf{z}_i' \mathbf{u}$ and $I_j \mathbf{z}_j' \mathbf{u}$ would be

$$Cov(I_i \mathbf{z}_i' \mathbf{u}, I_j \mathbf{z}_j' \mathbf{u}) = \pi_{ij} \left[\mathbf{Z}\boldsymbol{\Omega}\mathbf{Z}'\right]_{ij},$$

but we have so far ignored the inclusion probabilities in this case. We impose the following model on the sample:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}^*, Cov(\boldsymbol{\varepsilon}^*) = C^* = \sigma^2 diag(\pi_1, \pi_2, ..., \pi_n) .$$

In the ML-W and REML-W algorithms, certain matrix products $\mathbf{A}'\mathbf{B}$ ($\mathbf{A}, \mathbf{B} = \mathbf{X}, \mathbf{Z}, \mathbf{y}, \mathbf{e}$) in MSE are replaced by $\mathbf{A}'\mathbf{W}\mathbf{B}$, where $\mathbf{W}$ is the diagonal matrix of design weights. This is a sample-based estimate of the corresponding matrix product in population. In effect, we attempt to reconstruct the algorithm we would use if the population data were at hand. The resulting estimator is called EBLUPW. Alternatively, design information is incorporated into the model. For example, in PPS sampling, it is often a good idea to include the size variable in the model as an auxiliary variable (Lehtonen and Veijanen 2009).

Note that when the design weights are constant, REML and REML-W yield similar results. Then the domain estimates with the two methods should also be identical. The using of variable design weights of πPS has reduced the design bias in EBLUP estimation (Lehtonen et al., 2007).

3.1.3. Convergence problems

Especially with correlated area effects and correlated time effects, we have often encountered convergence problems. When the algorithm fails to converge, symptoms include very slow changes in estimated parameters, estimates converging to a boundary of the parameter space (for example, in time effect correlation, the correlation approaches 1), or parameters alternating between two or more values. The spatial correlation parameter may fluctuate wildly. In these cases, the end results are probably not useful. Then a simpler model with no correlations should be selected.

3.2. Synthetic estimator

The so-called synthetic estimator (SYN) is the sum of predictions over a domain:

$$\hat{t}_{dSYN} = \sum_{k \in U_d} \hat{y}_k . \qquad (16)$$

In survey methodology, the SYN estimator has been used with linear regression models. Variances of (16) are estimated by the diagonal elements of

$$\mathbf{X}_s Cov\left(\hat{\boldsymbol{\beta}}\right)\mathbf{X}_s',$$

where the subindex s refers to sums over population domains: The dth row of the matrix $\mathbf{X}_s$ is

$$\left[\mathbf{X}_s\right]_d = \sum_{i\in U_d}\left(x_{i1}, x_{i2}, ..., x_{ip}\right).$$

Synthetic estimator tends to have much smaller variance than GREG, but it often has large design bias. Therefore, the accuracy (measured by the MSE) and coverage properties of design-based confidence intervals can be poor.

Under a mixed model, a similar estimator to SYN is EBLUP(μ).

ANNEX 1. Domest installing instructions DRAFT

Domest needs Java JRE, version 6 or higher. SAS is required only if the population and sample input data sets are in SAS format.

Installing in a single computer: Domest is installed by executing DomestInstaller.jar. Choose the amount of memory available; probably best set to 1000, then click "I agree with conditions of use above", then select the folder where Domest is installed, and finally click Install.

Installing to several computers (a network environment): First, run DomestInstaller.jar in one computer and then copy the created folder with its contents and subfolders to the additional computers. The subfolders are needed by Domest. Domest is then executed in each computer by running Domest.jar or Domest.bat (on Windows).

ANNEX 2. Domest variables

Variables common to sample and population

Population should contain the variable determining the domains and all the x-variables used in the linear regression model or in the mixed model. Other required variables are summarized in the following table. All the variables must have identical names in the two data sets. A variable can be renamed on page 'Transformations'.

Variable	Requirements	Transform	When needed	Required in Population
y-variable	−	Indicator can be created after classification	Always	No, but if found, then true values appear in output
Design weights	Positive	−	In HT Hajek GREG pseudo-EBLUP REML-W	No
Domain variable	Usually integer	Classify	Always, unless only population totals and means required	Yes
x-variables	−	Indicators can be created after classification	Always, unless model has only indicator	Yes
Strata	Usually integer	−	In stratified sampling	Yes
Variable determining subsets for a random effect	Usually integer	Classify	In mixed models	Yes
Time variable	Usually integer	−	In mixed models with time effects or time-varying effects	Yes
Coordinate variables	−	−	In spatial mixed models	Yes

References

Berger, Y. G. (2004). A simple variance estimator for unequal probability sampling without replacement. *Journal of Applied Statistics* **31**, 305-315.

Berger, Y. G. (2005). Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics* **47**, 365-373.

Estevao, V. M., M. A. Hidiroglou, and C.-E. Särndal (1995). Methodological principles for a Generalized Estimation System at Statistics Canada. *Journal of Official Statistics* **11**, 181-204.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics* **35**, 1491-1523.

Hidiroglou, M. A. and Z. Patak (2004). Domain estimation using linear regression. *Survey Methodology* **30**, 67-78.

Lehtonen, R. and E. Pahkinen (2004). *Practical methods for design and analysis of complex surveys.* Second Edition. Wiley, Chichester.

Lehtonen, R., C.-E. Särndal, and A. Veijanen (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology* **29**, 33-44.

Lehtonen, R., C.-E. Särndal, and A. Veijanen (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition* **7**, 649-673.

Lehtonen, R., Myrskylä, M., C.-E. Särndal, and A. Veijanen (2007). Estimation for domains and small areas under unequal probability sampling. Invited paper at SAE 2007 Conference, Pisa, Italy (CD-ROM).

Lehtonen R. and Veijanen A. (2009). Design-based methods of estimation for domains and small areas. Chapter 31 in Rao C.R. and Pfeffermann D. (Eds.). *Handbook of Statistics. Sample Surveys: Inference and Analysis. Vol. 29B.* New York: Elsevier.

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

Rao, J. N. K. (2003). *Small area estimation*. Wiley: New York.

Saei, A. and R. Chambers (2004). Small area estimation under linear and generalized linear mixed models with time and area effects. EURAREA Consortium 2004, *Project Reference Volume*, www.statistics.gov.uk/eurarea.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model assisted survey sampling*. Springer-Verlag, New York.