EUROPEAN COMMISSION
Community Research

SEVENTH FRAMEWORK PROGRAMME

AMELI
Advanced Methodology for European Laeken Indicators

Deliverable 2.2

Small Area Estimation of Indicators on Poverty and Social Exclusion

Version: 2011

Risto Lehtonen, Ari Veijanen, Mikko Myrskylä
and Maria Valaste

The project FP7-SSH-2007-217322 AMELI is supported by European Commission funding from the Seventh Framework Programme for Research.

http://ameli.surveystatistics.net/

Contributors to Deliverable 2.2:

Chapter 1: Risto Lehtonen, Ari Veijanen, Mikko Myrskylä and Maria Valaste, University of Helsinki.

Chapter 2: Risto Lehtonen, Ari Veijanen, Mikko Myrskylä and Maria Valaste, University of Helsinki.

Chapter 3: Risto Lehtonen, Ari Veijanen, Mikko Myrskylä and Maria Valaste, University of Helsinki.

Chapter 4: Risto Lehtonen, Ari Veijanen, Mikko Myrskylä and Maria Valaste, University of Helsinki.

Chapter 5: Mikko Myrskylä, University of Helsinki.

Chapter 6: Risto Lehtonen, Ari Veijanen, Mikko Myrskylä and Maria Valaste, University of Helsinki.

Main responsibility

Risto Lehtonen, University of Helsinki

Data provision and commenting

Timo Alanko, Pauli Ollila, Marjo Pyy-Martikainen, Statistics Finland; Rudi Seljak, Statistics Slovenia; Kaja Sõstra, Statistics Estonia.

Evaluators

Internal evaluator: Matthias Templ, Vienna University of Technology.

Aim and Objectives of Deliverable 2.2

There is increasing user demand for regional or sub-population official statistics within the EU. In many countries, statistics on poverty and social exclusion are based on sample surveys, such as the SILC survey. One of the aims stated for the AMELI project was to investigate the adaptation of modern small area and domain estimation (SAE) approaches for selected indicators on poverty and social exclusion (Laeken indicators). At-risk-of poverty rate, the Gini coefficient, relative median at-risk-of poverty gap and quintile share ratio were selected for consideration. Estimation approaches examined in Work Package 2 involved the use of auxiliary population data and statistical models for borrowing strength for regional (e.g. area sizes below NUTS3) and small area estimation purposes. The methods included design-based model-assisted estimators and model-based estimators. The relative merits and practical applicability of the methods was assessed by simulation experiments using real register and survey data. It was considered important to cover a broad variety of typical practical estimation settings existing in different EU countries. Therefore, the methods were investigated under various statistical infrastructures, sampling designs, domain compositions and outlier contamination schemes. In many cases, the methods assumed access to unit-level auxiliary population data. This option is becoming increasingly realistic in statistical infrastructures of the EU countries, where opportunities to use administrative registers and population census data for statistical purposes are improving. Methods were also developed that use aggregate-level auxiliary data, which option is useful for countries where aggregate auxiliary data are available for example from official statistics sources. The accompanying R programs codes were provided for practical application of the methods. In the production of Deliverable 2.2 on small area statistics methodology, the aim was to combine expertise from academic research with expertise from Official statistics producers. NSIs involved include Statistics Finland, Statistics Estonia and Statistics Slovenia. University of Helsinki has the main responsibility of the production of the deliverable.

Contents

1 Introduction 1

1.1 Objectives 1

1.2 Basic approaches 2

1.2.1 Estimation approaches 2

1.2.2 Report structure 4

1.3 Planned and unplanned domain structures 5

1.4 Direct and indirect estimators 6

1.5 Estimation of poverty indicators 6

1.6 The role of models and auxiliary data 8

1.6.1 The role of models 8

1.6.2 The role of auxiliary information 9

1.6.3 Estimation under outlier contamination 11

2 Basic properties of domain estimators 12

3 Models and estimators 13

3.1 Models and auxiliary data 14

3.2 Design-based estimators 17

3.2.1 Horvitz-Thompson estimator 17

3.2.2 Generalized regression estimator 17

3.2.3 Model calibration 18

3.3 Model-based estimators 23

3.3.1 Synthetic estimator 23

3.3.2 EBLUP and EBP estimators 23

3.4 Transformations of predictions 25

3.5 Frequency-calibrated predictors calculated using known domain marginal totals of auxiliary variables 29

3.6 Composite estimators 33

3.7 Simulation-based methods 36

4 Estimators for poverty indicators and results of Monte Carlo simulation experiments 38
4.1 Introduction 38
4.2 Experimental design 38
4.2.1 Register-based population from Western Finland 38
4.2.2 Amelia population 41
4.2.3 Quality measures 41
4.2.4 Contamination schemes 42
4.2.5 Estimators 43
4.3 At-risk-of poverty rate 44
4.3.1 HT-CDF estimator 45
4.3.2 Methods based on poverty indicators 45
4.3.3 Simulation results 47
4.4 The Gini coefficient 53
4.5 Poverty gap 58
4.6 Quintile share ratio S20/S80 63
4.7 Classifying domains by poverty 69

5 Case study: Estimation of poverty rate and its variance 71
5.1 Introduction 71
5.2 Design 71
5.3 Estimators 74
5.3.1 Poverty rate estimators 74
5.3.2 Variance estimators 75
5.4 Results 78
5.4.1 Poverty rate estimators 78
5.4.2 Variance estimators 80

6 Discussion of results 82
6.1 General 82
6.2 New predictors 82
6.3 Comparison of outlier and contamination mechanisms 84

References 86

Annex 1. Manual for R codes 93

Annex 2. AMELI WP 2 Estimation: Summary of SAE methods 98

Annex 3. Technical summary of selected estimator types 100

1 Introduction

1.1 Objectives

There are increasing needs in the society for accurate statistics on poverty and social exclusion (poverty indicators for short) produced for different population subgroups or domains such as regional areas and demographic groups. One of the aims of the AMELI project was to investigate the current (standard) methods for domain and small area estimation of poverty indicators and develop new methods where appropriate. This report presents the methodological developments and summarizes our main findings on statistical properties of proposed estimators.

Properties of estimators of selected poverty indicators (so-called Laeken indicators as agreed in Laeken European Council in December 2001) were studied by simulation experiments. The study had the following objectives:

1. Investigation of statistical properties (bias and accuracy) of standard direct estimators of the selected poverty indicators for population domains and small areas. Standard estimators do not use auxiliary data or modelling.
2. Introduction of alternative estimators, which use statistical models and auxiliary data at the unit level, and investigation of bias and accuracy of the new estimators.
3. Introduction of estimators that use auxiliary data at an aggregated level and investigation of bias and accuracy of these estimators.
4. Implementation of points 1 to 3 under equal and unequal probability sampling schemes.
5. For studying robustness of methods, the implementation of points 1 to 4 under various outlier contamination schemes.
6. Study of applicability of a method incorporating a novel transformation of predictions.
7. Implementation of points 1 to 5 for populations from two different data sources, register-based data maintained by Statistics Finland (the Western Finland population) and sample survey data from EU-wide SILC survey (the Amelia population).

1.2 Basic approaches

1.2.1 Estimation approaches

This report presents the research done at University of Helsinki in the context of AMELI Work Package 2 on the estimation of selected indicators on poverty (monetary Laeken indicators) for domains and small areas. Domain estimation of poverty has been recently studied by D'Alo et al. (2006), Fabrizi et al. (2007a, 2007b), Srivastava (2009), Molina and Rao (2010), and Haslett et al. (2010). Verma et al. (2010) reports empirical results for regional estimation using EU-SILC data.

The indicators considered in this report are the following:

- At-risk-of poverty rate
- The Gini coefficient
- Relative median at-risk-of poverty gap
- Quintile share ratio (S20/S80 ratio).

The indicators are typically nonlinear and are based on non-smooth functions such as medians and quintiles, which makes the estimation a non-trivial task. This holds especially for the estimation of the indicators for domains and small areas.

In this report, both design-based and model-based or model-dependent methods are developed and investigated for the estimation of the selected poverty indicators for domains and small areas. Design-based methods are chosen because of the dominance of the framework in official statistics production. Model-based approaches are important to be covered because in many small area estimation situations, model-based methods provide a realistic solution.

Design-based estimation for finite population parameters refers to an estimation approach where the randomness is introduced by the sampling design. In design-based estimation, it is emphasized that estimators should be design consistent and, preferably, nearly design unbiased at least in domains with medium-sized samples (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 (Särndal, 2007, p. 99). This property is independent of the choice of the assisting model. Generalized regression (GREG) type estimators and calibration type estimators are examples of nearly design unbiased estimators. Model-assisted GREG estimators are constructed such that they are robust against model mis-specification.

GREG and model-free calibration are discussed in Särndal, Swensson and Wretman (1992) and Särndal (2007). Lehtonen and Veijanen (2009) discuss GREG and model-free calibration in the context of domain estimation. In calibration, we concentrate on model calibration estimators, introduced in Wu and Sitter (2001). Model calibration has been developed for domain estimation in Lehtonen, Särndal and Veijanen (2009). In GREG and model calibration we often employ estimators that use nonlinear assisting models involving random effects in addition to the fixed effects.

Design-based estimators for domains and small areas are usually constructed so that the complexities of the sampling design, such as stratification and unequal inclusion probabilities, are accounted for. For example, it is customary that design weights are incorporated in a design-based estimation procedure. This does not necessarily hold for model-based or model-dependent methods. In this respect, a conceptual separation of model-based and model-dependent methods can be helpful. In strict model-dependent methods, the estimation is considered to rely exclusively on the statistical model adopted. For example, design weights do not play any role in a model-dependent estimation procedure. For design consistency, variables that capture (at least some) of the sampling complexities, such as stratification variables and PPS size variable, can be included in the underlying model. In model-based methods, design weights can be incorporated in the estimation procedure to account for unequal probability sampling, leading to design consistent pseudo synthetic, pseudo EBLUP (empirical best linear unbiased predictor) and pseudo EBP (empirical best predictor) type approaches (see e.g. Rao, 2003; You and Rao, 2002; Jiang and Lahiri, 2006). The methods coincide under equal probability sampling. In this report, we use "model-

based" as a general concept unless it is instructive to treat separately the two approaches.

Model-based estimators can have desirable properties under the model but their design bias does not necessarily tend to zero with increasing domain sample size (Hansen, Hurvitz and Madow, 1978; Hansen, Madow, and Tepping, 1983; Särndal, 1984, and Lehtonen, Särndal and Veijanen, 2003). Model-based methods for small area estimation include a variety of techniques such as synthetic (SYN) and composite estimators, EBLUP and EBP type estimators and various Bayesian techniques, such as empirical Bayes and hierarchical Bayes. The monograph by J.N.K. Rao (2003) provides a comprehensive treatment of model-based small area estimation (SAE). Mixed models that are commonly used in SAE are discussed for example in Jiang and Lahiri (2006).

Model-based small area estimation methodology was extensively studied in the context of the EU's FP6 research project EURAREA (Enhancing Small Area Estimation Techniques to meet European Needs, 2002-2004), see The EURAREA Consortium (2004). EURAREA concentrated mainly on the estimation of small area totals and means and recommended the model-based methods for official statistics production for small areas (e.g. area sizes below NUTS3). In AMELI we extend the SAE methodology to considerably more complex statistics including the Gini coefficient, relative median at-risk-of poverty gap and quintile share ratio. In addition to model-based methods, advanced design-based methods are developed.

1.2.2 Report structure

The report includes the description of the estimators developed for the selected poverty indicators and the results of the Monte Carlo simulation experiments on the statistical properties (bias and accuracy) of the estimators. The report is organized as follows. The remainder of this section covers the definition of the basic concepts and introduces the estimators of the poverty indicators to be examined as well as the role of models and auxiliary data in the construction of the estimators. Section 2 summarizes the basic properties of the various estimator types for domains and small areas. A technical description of the models and estimators is inserted in Section 3.

Section 4 contains a detailed description of the specific estimators of the indicators and presents the results of Monte Carlo experiments. Section 5 is devoted to a case study on a model-assisted estimator of poverty rate; special attention is in the estimation of the variance of the estimator. Discussion is in Section 6.

1.3 Planned and unplanned domain structures

Different domain structures can appear in practical applications of domain estimation (Lehtonen and Veijanen, 2009). 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, Gambino and Mantel, 1994). For planned domain structures, the population domains can be regarded as separate subpopulations. Therefore, standard population estimators are applicable as such. The domain size in every domain is often assumed known and the sample size n_d in domain sample s_d is fixed in advance. Stratified sampling in connection to a suitable allocation scheme such as optimal (Neyman) or power (Bankier) allocation is often used in practical applications, in order to obtain control over domain sample sizes (e.g. Lehtonen and Pahkinen, 2004). Singh, Gambino and Mantel (1994) describe allocation strategies to attain reasonable accuracy for small domains, still retaining good accuracy for large domains. Falorsi, Orsini and Righi (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 will be random. The domains are then called *unplanned*. 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. Unplanned domain structures are often assumed in this report.

1.4 Direct and indirect estimators

It is advisable to separate direct and indirect estimators for domains (Lehtonen and Veijanen, 2009). 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 Horvitz-Thompson (HT) type estimator provides a simple example of direct estimator. In model-assisted estimation, direct estimators are constructed by using models fitted separately in each domain. 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). In general, indirect estimators are attempting to "borrow strength" from other domains and/or in a temporal dimension. Indirect model-assisted estimators for domains are discussed in the literature (e.g. Estevao and Särndal, 1999, Lehtonen, Särndal and Veijanen, 2003, 2005, and Hidiroglou and Patak, 2004). Indirect estimators are used extensively in this report; this especially holds for domains whose sample size is small. Direct estimators are occasionally used in cases where the domain sample sizes are large. Direct estimators also serve as reference or benchmark estimators when investigating the bias and accuracy of the proposed indirect estimators.

1.5 Estimation of poverty indicators

The poverty (Laeken) indicators discussed in this report can be divided into two groups with respect to the selected estimation approach. For the estimation of at-risk-of poverty rate based on poverty indicators, we use GREG and model calibration type estimators (featuring design-based model assisted methods) and SYN and EBLUP or EBP type estimators (featuring model-based methods). In all these estimators, logistic models are used because the underlying study variable is binary. Direct estimators,

such as Horvitz-Thompson type estimators, are used as basic or reference estimators, sometimes also called “default” estimators in this report.

In addition to the estimation of poverty rate for domains and small areas, we have examined methods for the identification of domains that can be characterized as “poor”, i.e. domains whose estimated poverty rate falls below a given threshold. Ranking of domains is part of so-called *triple-goal estimation*, where the goal is to obtain good ranks, good histogram and accurate domain estimates (Rao, 2003; Shen and Louis, 1998; Paddock et al., 2006). Judkins and Liu (2000) present methods for improving the estimated range of domain estimators.

The equivalized income constitutes the key variable underlying the poverty (monetary Laeken) indicators. Equivalised income is defined as the household's total disposable income divided by its "equivalent size", to take account of the size and composition of the household, and is attributed to each household member (including children) (European Commission, 2006). Equivalization is made on the basis of the OECD modified scale, which assigns weight 1.0 for the first adult, 0.5 for every additional person aged 14 or over, and 0.3 for every child under 14. Relative median at-risk-of poverty gap (poverty gap for short) and quintile share ratio (S20/S80 ratio) are examples of indicators that rely on medians or quantiles of the cumulative distribution function (CDF) of the underlying continuous variable. For these indicators, HT type direct estimators, synthetic and composite estimators are developed. A composite estimator is constructed as a linear combination of a design-based direct estimator and a model-based SYN estimator. In addition, for poverty gap we have studied estimation of conditional expectations by simulation-based methods, resembling methods introduced in Molina and Rao (2010). In constructing the estimators, we use logarithmic transformation to correct for the skewness of the distribution of the study variable. In back-transformation we first tried the RAST (Ratio Adjusted by Sample Total; Chambers and Dorfman, 2003, Fabrizi et al., 2007b) type transformation, and later developed more elaborate transformations aimed at improving the histogram of transformed predictions.

The statistical properties (design bias and accuracy) of the estimators of the selected poverty indicators are examined with Monte Carlo simulation experiments. Real data

taken from statistical registers of Statistics Finland are used in constructing the frame populations. We have made experiments also with the synthetic Amelia population (Alfons et al. 2011b). The populations contain a wide selection of socio-economic and demographic auxiliary variables. We have concentrated on design-based simulation settings.

Programs written in R language have been produced for statistical computing of the selected poverty indicators for domains and small areas. The R codes are described in a separate supplemental deliverable Veijanen and Lehtonen (2011).

1.6 The role of models and auxiliary data

1.6.1 The role of models

Choice of statistical model underlying an estimator of a poverty indicator constitutes an important phase of the estimation procedure for domains and small areas. In constructing model-assisted and model-based estimators, we use selected models from the family of generalized linear mixed models (GLMM, e.g. McCulloch and Searle, 2003). Linear and logistic fixed-effects and mixed models are extensively used. Lehtonen, Särndal and Veijanen (2003, 2005) discuss the choice of the model in the context of GREG estimation.

The rationale behind the choice of the assisting model for GREG is the following. In GREG estimation for domains, various types of study variables can be used. For example, a linear model formulation is appropriate for a continuous variable, and logistic models are usually chosen for binary or polytomous variables. We call “extended GREG family” the GREG estimators that use GLMM’s as assisting models.

In the parametrization of the assisting model for an extended GREG family estimator, it is important for accurate domain estimation to account for the possible domain differences. Basically, domain differences can be accounted for either with a fixed-

effects or a mixed model specification. A fixed-effects model is usually a default in GREG estimation. Mixed model specification offers a flexible approach for domain estimation (Lehtonen, Särndal and Veijanen, 2003, 2005) and is much used in our research. Because of this model choice, the resulting estimators for domains are in most cases of indirect type.

1.6.2 The role of auxiliary information

The availability of high-quality auxiliary information is crucial for reliable estimation for domains and small areas. Auxiliary information can be incorporated into the sampling design (e.g. stratified sampling, PPS sampling) or into the estimation procedure (or both). Stratified sampling is often used to obtain sufficient sample size for the most important domains of interest (leading to planned domains). In this report we concentrate on the use of auxiliary data in the estimation procedure. Both equal probability and unequal probability sampling design are discussed, under unplanned domain structures (referring to cases where the domains of interest are not defined as strata in the sampling design).

The reason for incorporating auxiliary data in an 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 estimation for domains and small areas. The auxiliary data can be aggregated at the population level or 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 in many European countries and 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.

A rapidly progressing trend in official statistics production is the use of unit-level auxiliary data for domain and small area estimation. These data are incorporated in the estimation procedure by unit-level statistical models. Under this option, register data (such as population census register, different unit-level administrative and statistical registers) can be available as frame populations and sources of auxiliary

data. Moreover, the registers often contain unique identification keys that can be used in merging at micro level different register sources and data from registers and sample surveys. Known domain membership for all population elements is often assumed. Many countries, both in Europe and in the European Union, are progressing in the development of reliable population registers that can be accessed for statistical purposes. Good examples are Austria, Estonia, Finland and Slovenia, which have representation in the AMELI project. Obviously, access to micro-merged register and survey data provides great flexibility for the development of methods for domain estimation and in the domain estimation practice.

All estimator types (except HT and related direct estimators) examined in this report aim at using information about auxiliary variables in the population. We have first assumed access to unit-level auxiliary information. The reason is that this option offers much flexibility for estimator construction. Under this option, a model is fitted to the sample data, predictions are calculated for all population elements using the estimated model parameters and the known values of the auxiliary variables, and the predictions in the population contribute to the estimation of the indicators of interest, such as poverty rate in the given domains and small areas.

Because the option of the use of unit level auxiliary data for statistical purposes is not (yet) commonly available in statistical infrastructures within the EU, we extend the methodology to cases where only aggregate-level auxiliary data are available. In the method we only assume that the population totals of continuous auxiliary variables, or population frequencies of classes of discrete variables, are known. A calibration method is introduced to calculate the necessary predicted values.

We have not applied Bayesian methods (e.g., Fabrizi et al., 2005) or models involving spatial or temporal correlations (Chandra et al., 2007). SAE methods that borrow strength in spatial or temporal dimension were developed and investigated to some extent in the context of the EU's FP5 project EURAREA.

1.6.3 Estimation under outlier contamination

In developing estimators that are robust against outlier contamination we discuss the contamination mechanisms and models proposed in the WP4 working document by Hulliger and Schoch (2010). Outlying mechanisms considered are OCAR (outlying completely at random) and OAR (outlying at random), and the contamination models are CCAR (contaminated completely at random), CAR (contaminated at random), and NCAR (not contaminated at random). The definitions of these concepts are given in the working document referred above.

2 Basic properties of domain estimators

Known design-based properties related to bias and accuracy of design-based model-assisted estimators and model-dependent estimators for domains and small areas are summarized in Table 1 (Lehtonen and Veijanen, 2009). Model-assisted estimators such as GREG and calibration 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 mean squared error (MSE), as shown for example by Lehtonen, Särndal and Veijanen (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 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 minor or 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 mini domain when its share of population is smaller than 1% . In so small domains, especially direct estimators can have large variance. Small domains are the main reason to prefer indirect model-based estimators to direct design-based estimators (Rao, 2003).

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

	Design-based model-assisted methods GREG and calibration estimators	**Model-based and model-dependent methods** 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 (Mean Squared Error, 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

In practice, there are two main approaches to design-based estimation for domains: direct estimators that are usually applied for planned domain structures (such as strata whose sample sizes n_d are fixed in the sampling design) and indirect estimators whose natural applications are for unplanned domains (whose domain sample sizes are random). In model-based or model-dependent SAE, indirect estimators that aim at "borrowing strength" are often used.

3 Models and estimators

The fixed and finite population of interest is denoted $U = \{1, 2, ..., k, ..., N\}$, where k refers to the label of population element. A *domain* is a subset of population U such as a regional population in NUTS3 or NUTS4 region or a demographic subdivision within the regional areas. Poverty rate estimates, for example, are required not only for regions but also for classes defined by age and gender. Consider a region r and a class c. They define a domain d: in population U, a subset $U_d = U_r \cap U_c$ contains people belonging to class c (U_c) in region r (U_r). The number of units in the domain in population is denoted by N_d. In sample s, corresponding subsets are defined as $s_d = s_r \cap s_c$ with n_d observations. Naturally, regions are special cases of domains. A small area is a domain whose realized sample size is small (even zero).

Many poverty indicators are composed of domain totals, frequencies and medians. The domain total of the study variable y (equivalized incomes) is defined as

$$t_d = \sum_{k \in U_d} y_k \,, \tag{1}$$

where y_k denotes the value of the study variable for element k. The frequency f_d of a class C, such as the frequency of persons with income smaller than a threshold, is written as a sum of class indicators $v_k = I\{y_k \in C\}$:

$$f_d = \sum_{k \in U_d} v_k \,. \tag{2}$$

For a binary indicator, (1) and (2) obviously coincide.

3.1 Models and auxiliary data

Auxiliary information is used in model-assisted and model-based methods. The available auxiliary information consists of an auxiliary x-vector and a domain membership specification $I_{dk} = 1$ if $k \in U_d$, $I_{dk} = 0$ otherwise, $d = 1,...,D$, for every unit $k \in U$. Letting $\mathbf{x}_k$ denote the value of the auxiliary vector for unit *k*, we thus assume that both $\mathbf{x}_k$ and domain membership I_{dk} is known for every $k \in U$.

Models are incorporated in model-assisted (GREG, model calibration) and model-based (synthetic, EBLUP, EBP) methods. Consider 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 (Lehtonen and Veijanen, 2009). Examples of $f(\cdot;\boldsymbol{\beta})$ are a linear functional form and a logistic function. The model fit to the sample data $\{(y_k, \mathbf{x}_k);\ k \in s\}$ yields the estimate $\hat{\boldsymbol{\beta}}$ 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 $\hat{y}_k = f(\mathbf{x}_k;\hat{\boldsymbol{\beta}})$ for every $k \in U$, which is possible under our assumptions.

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

Let us discuss linear models in more detail. For a linear fixed-effects model

$$Y_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k, k \in U$$

we derive two special cases, a common model formulation and a model formulation involving domain-specific intercepts.

Under the common model formulation, we have $\mathbf{x}_k = (1, x_{1k}, ..., x_{Jk})'$, known for every $k \in U$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_J)'$ where β_j are fixed effects common for all domains, $j = 0, ..., J$. Under the model formulation with domain-specific intercepts, we have $\mathbf{x}_k = (I_{1k}, ..., I_{Dk}, x_{1k}, ..., x_{Jk})'$, $I_{dk} = 1$ if $k \in U_d$, $I_{dk} = 0$ otherwise, $d = 1, ..., D$, and $\boldsymbol{\beta} = (\beta_{01}, ..., \beta_{0D}, \beta_1, ..., \beta_J)'$, where β_{0d} are domain-specific intercepts and β_j are common slopes, $j = 1, ..., J$. In both special cases, predicted values $\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}}$ are calculated for every $k \in U$.

The rationale behind the two special cases is the following. If a single (common) fixed-effects model is assumed to hold in all domains, possible differences between domains are not necessarily captured by the estimator, although in GREG the weighted sum of residuals corrects for design bias caused by the possible model misspecification. For fixed effects model, there is some theoretical support for using domain-specific intercepts, or at least regional indicators, to account for possible differences between regions. Then the beta parameters, or slopes, associated with explanatory x-variables are often specified common to all domains. The two special cases of models result in an indirect domain estimator.

A direct estimator is obtained by using separate slopes for every domain in addition to the separate intercepts, that is, a model $Y_k = \mathbf{x}_k' \boldsymbol{\beta}_d + \varepsilon_k, k \in U_d$. This model would probably result in too unstable domain estimates, in particular if the domain sample size is small. On the other hand, a domain-specific model might be realistic for domains with a large sample size.

In order to account for possible differences between regions, a linear mixed model incorporates domain-specific random effects $u_d \sim N(0, \sigma_u^2)$ for domain U_d, or regional random effects $u_r \sim N(0, \sigma_u^2)$ for region U_r, where $U_d \subset U_r$. For domain-specific random intercepts, a linear mixed model is given by

$$Y_k = \mathbf{x}_k' \boldsymbol{\beta} + u_d + \varepsilon_k, k \in U_d, \varepsilon_k \sim N(0, \sigma^2),$$

or, more generally,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \varepsilon$$

for a matrix $\mathbf{Z}$. The parameters $\boldsymbol{\beta}$, σ_u^2 and σ^2 are first estimated from the data, and the values of the random effects are then predicted.

An example of a generalized linear mixed model formulation is a binomial logistic mixed model for a binary *y*-variable. We want to estimate the totals $t_d = \sum_{k \in U_d} y_k$ for all domains U_d. The logistic mixed model is of the form

$$E_m(y_k \mid \mathbf{u}_d) = P\{y_k = 1 \mid \mathbf{u}_d\} = \frac{\exp(\mathbf{x}_k'(\boldsymbol{\beta} + \mathbf{u}_d))}{1 + \exp(\mathbf{x}_k'(\boldsymbol{\beta} + \mathbf{u}_d)}$$

for $k \in U_d$, $d = 1,...,D$, where $\mathbf{x}_k$ is a known vector value for every $k \in U$, $\boldsymbol{\beta}$ is a vector of fixed effects common for all domains, and $\mathbf{u}_d$ is a vector of domain-specific random effects. Here again, predictions

$$\hat{y}_k = \exp(\mathbf{x}_k'(\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_d)) / (1 + \exp(\mathbf{x}_k'(\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_d))$$

are calculated for every $k \in U$. Lehtonen, Särndal and Veijanen (2005) give several special cases of the model. An indirect estimator for domains is obtained with mixed model specification.

We have fitted most of the mixed models with R function nlme. By default it uses the maximum likelihood method. In nlme, the design weights do not contribute to estimation. Design weights can be included in model fitting with R function glmer (package lme4). When fitting the fixed effects models, we have used design weights.

3.2 Design-based estimators

3.2.1 Horvitz-Thompson estimator

Horvitz-Thompson (HT) estimator of domain total (1) is a weighted sum of values in the sample:

$$\hat{t}_d = \sum_{k \in s_d} a_k y_k \,, \tag{3}$$

where the design weights a_k are inverses of inclusion probabilities π_k ($a_k = 1/\pi_k$). An HT estimator is a direct estimator. It does not incorporate any model. The estimator is design unbiased but it can have large variance, especially for small domains. HT estimator is often used under planned domain structures, where the domain sample sizes are sufficiently large.

3.2.2 Generalized regression estimators

Generalized regression (GREG) estimators (Särndal et al., 1992; Lehtonen and Veijanen, 2009) are assisted by a model fitted to the sample. By choosing different models we obtain a family of GREG estimators with same form but different predicted values (Lehtonen et al., 2003, 2005, 2007).

Ordinary GREG estimator

$$\hat{t}_{d;GREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k) \tag{4}$$

incorporating a linear regression model is used to estimate domain totals (1) of a continuous study variable. For a binary or polytomous response variable, a linear model formulation will not necessarily fit the data well. A logistic model formulation might be a more realistic choice. LGREG (logistic GREG; Lehtonen and Veijanen, 1998) estimates the frequency f_d of a class *C* in each domain. A logistic regression

model is fitted to the indicators $v_k = I\{y_k \in C\}$, $k \in s$, using the design weights. The fitted model yields estimated probabilities $\hat{p}_k = P\{v_k = 1; \mathbf{x}_k, \hat{\boldsymbol{\beta}}\}$. The LGREG estimator of the class frequency in U_d is

$$\hat{f}_{d;LGREG} = \sum_{k \in U_d} \hat{p}_k + \sum_{k \in s_d} a_k (v_k - \hat{p}_k). \qquad (5)$$

Here $\sum_{k \in U_d} \hat{p}_k$ is the sum of predicted values in the population. Thus it is necessary to have access to unit level population information about the persons' auxiliary variables. The last component of (5), i.e. an HT estimator of the residual total, aims at correcting the possible bias of the first (synthetic) part. It is obvious that for certain model choices, notably for a domain-specific model formulation, the last component vanishes.

A so-called domain size correction (Lehtonen and Veijanen, 2009) is incorporated into an estimator defined as

$$\hat{f}_{d;LGREG(2)} = \sum_{k \in U_d} \hat{p}_k + \frac{N_d}{\hat{N}_d} \sum_{k \in s_d} a_k (v_k - \hat{p}_k);\ \hat{N}_d = \sum_{k \in s_d} a_k . \qquad (6)$$

In the MLGREG estimator (Lehtonen and Veijanen, 1999, Lehtonen, Särndal and Veijanen, 2005, Torabi and Rao, 2008), an alternative logistic mixed model involving fitted values $\hat{p}_k = P\{v_k = 1; \mathbf{x}_k, \hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}_d\}$ is used instead of a fixed-effects logistic model. The random effects are associated with domains or with regions. This model formulation may be a realistic option for many situations in practice.

3.2.3 Model calibration

Calibration is typically used to construct an estimator as weighted sample sum with weights chosen so that the weighted sample sums of auxiliary variables are identical with known population totals (Estevao and Särndal, 2004; Kott, 2009). In *model calibration* introduced by Wu and Sitter (2001) and Wu (2003), predictions are used instead of auxiliary variables. We have generalized model calibration for domain

estimation (Lehtonen et al., 2009). A model is first fitted to the sample. We discuss only a logistic regression model, although any model could be applied. The estimator of the total frequency is a weighted sum of indicators over the whole sample, region or the domain. The weights are chosen so that the weighted sum of estimated probabilities over a subset of sample equals the sum of predicted probabilities over a corresponding subset of population. The sum of weights over the sample subset must equal the size of the population subset. Moreover, the weights should be close to the design weights. The procedure of finding such weights is called calibration (e.g. Särndal, 2007).

In population level calibration (Wu and Sitter, 2001), the weights must satisfy calibration equation

$$\sum_{i\in s} w_i z_i = \sum_{i\in U} z_i = \left(N, \sum_{i\in U} \hat{p}_i \right), \qquad (7)$$

where $z_i = (1, \hat{p}_i)$. Using the technique of Lagrange multiplier (λ), we minimize

$$\sum_{k\in s} \frac{(w_k - a_k)^2}{a_k} - \lambda' \left(\sum_{i\in s} w_i z_i - \sum_{i\in U} z_i \right)$$

under the conditions (7). The first part of the equation is the distance between the weights w_k and the known design weights a_k. The latter part corresponds to the constraints (7). The equation is minimized by weights

$$w_k = a_k g_k; \quad g_k = 1 + \lambda' z_k,$$

where

$$\lambda = \left(\sum_{i\in U} z_i - \sum_{i\in s} a_i z_i \right)' \left(\sum_{i\in s} a_i z_i z_i' \right)^{-1}.$$

The domain estimator is defined as a domain sum

$$\hat{f}_{d;pop} = \sum_{k \in s_d} w_k v_k \ . \tag{8}$$

In our experiments, this estimator has not performed well.

The first choice for domain level calibration is equation

$$\sum_{i \in s_d} w_{di} z_i = \sum_{i \in U_d} z_i = \left(N_d, \sum_{i \in U_d} \hat{p}_i \right), \tag{9}$$

where the weights w_{di} are specific to the domain. From (9) we see that the domain sizes must be known. We minimize

$$\sum_{k \in s_d} \frac{(w_{dk} - a_k)^2}{a_k} - \lambda_d' \left(\sum_{i \in s_d} w_{di} z_i - \sum_{i \in U_d} z_i \right)$$

under the calibration equations (9). The solution is

$$w_{dk} = a_k g_{dk}; \quad g_{dk} = 1 + \lambda_d' z_k ,$$

where

$$\lambda_d = \left(\sum_{i \in U_d} z_i - \sum_{i \in s_d} a_i z_i \right)' \left(\sum_{i \in s_d} a_i z_i z_i' \right)^{-1} .$$

The frequency in the domain is estimated by a weighted sum of indicators over the domain:

$$\hat{f}_{d;s} = \sum_{k \in s_d} w_{dk} v_k \ . \tag{10}$$

We call this estimator semi-direct, referring to the fact that the sum contains only observations from the domain. It is not direct, however, as the weights are determined by a fitted model that incorporates all sample values. Next we introduce some semi-indirect estimators incorporating observations outside the domain.

The first semi-indirect domain level calibration estimator is a sum over the whole sample with domain-specific weights w_{dk} that are close to weights a_k in the domain and close to zero outside the domain. In other words, the weights should be close to $I\{k \in s_d\}a_k = I_{dk}a_k$ ($I_{dk} = I\{k \in s_d\}$). The calibration equation is

$$\sum_{i \in s} w_{di} z_i = \sum_{i \in U_d} z_i \,. \tag{11}$$

We minimize

$$\sum_{k \in s} \frac{(w_{dk} - I_{dk}a_k)^2}{a_k} - \lambda_d' \left(\sum_{i \in s} w_{di} z_i - \sum_{i \in U_d} z_i \right).$$

The solution is

$$w_{dk} = I_{dk}a_k + \lambda_d' a_k z_k \,;$$

$$\lambda_d = \left(\sum_{i \in U_d} z_i - \sum_{i \in s} I_{di} a_i z_i \right)' \left(\sum_{i \in s} a_i z_i z_i' \right)^{-1} .$$

The estimator is defined as a weighted sum over the whole sample:

$$\hat{f}_{d;s} = \sum_{k \in s} w_{dk} v_k \,. \tag{12}$$

Alternatively, the summation extends only over the domain.

We have also considered a similar estimator defined as a regional sum:

$$\hat{f}_{d;s} = \sum_{k \in s_r} w_{dk} v_k \ , \qquad (13)$$

where the subset s_r of sample contains all the people in the same region *r* as the domain. The calibration equation is

$$\sum_{i \in s_r} w_{di} z_i = \sum_{i \in U_d} z_i \ .$$

We minimize

$$\sum_{k \in s_r} \frac{(w_{dk} - I_{dk} a_k)^2}{a_k} - \lambda' \left(\sum_{i \in s_r} w_{di} z_i - \sum_{i \in U_d} z_i \right)$$

obtaining

$$w_{dk} = I_{dk} a_k + \lambda_d' a_k z_k \ ;$$

$$\lambda_d = \left(\sum_{i \in U_d} z_i - \sum_{i \in s_r} I_{di} a_i z_i \right)' \left(\sum_{i \in s_r} a_i z_i z_i' \right)^{-1} \ .$$

This estimator apparently ”borrows strength” from other domains in same region. Estevao and Särndal (2004) have shown that borrowing strength is not always a good idea, but they consider a different class of calibration estimators. In contrast with their estimators, our estimator is a sum over a set larger than the domain, and the weights are close to zero outside the domain.

3.3 Model-based estimators

3.3.1 Synthetic estimator

Synthetic (SYN) estimator is typically a sum of predicted values over the population elements in a domain. In the case of a logistic model, synthetic estimator is the sum of predicted probabilities:

$$\hat{f}_{d;LSYN} = \sum_{k \in U_d} \hat{p}_k \,. \tag{14}$$

For logistic SYN (LSYN) estimator using a logistic fixed-effects model, the predictions are $\hat{p}_k = P\{v_k = 1; \mathbf{x}_k, \hat{\boldsymbol{\beta}}\}$, and $\hat{p}_k = P\{v_k = 1; \mathbf{x}_k, \hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}_d\}$ for a MLSYN estimator using a logistic mixed model. Obviously, LSYN estimator (14) constitutes the first component of the LGREG estimator (5).

3.3.2 EBLUP and EBP estimators

The *EBLUP estimator* (empirical best linear unbiased estimator, e.g. Rao, 2003, p. 95) is used in the context of a linear mixed model

$$Y_k = \mathbf{x}_k'\boldsymbol{\beta} + u_d + \varepsilon_k, k \in U_d \,,$$

or, more generally,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \varepsilon$$

for a matrix $\mathbf{Z}$. Under the first mixed model the domain total's conditional expectation given the random effects $\mathbf{u}$ is

$$E\left(\sum_{k \in U_d} Y_k \,\middle|\, \mathbf{u}\right) = \left(\sum_{k \in U_d} \mathbf{x}_k\right)' \boldsymbol{\beta} + N_d u_d \,.$$

This would be an optimal predictor of the domain total in the sense of minimizing MSE. Its best linear unbiased predictor (BLUP) is

$$\hat{t}_{BLUP} = \left(\sum_{k \in U_d} \mathbf{x}_k \right)' \hat{\boldsymbol{\beta}}(\sigma_u^2, \sigma^2) + N_d \hat{u}_d(\sigma_u^2, \sigma^2),$$

where the optimal estimators of $\boldsymbol{\beta}$ and $\mathbf{u}$ depend on unknown variance components σ_u^2 and σ^2 as follows: For $\mathbf{R} = Cov(\varepsilon; \sigma^2)$, $\mathbf{G} = Cov(\mathbf{u}; \sigma_u^2)$ and $\mathbf{V} = \mathbf{R} + \mathbf{ZGZ}'$,

$$\hat{\boldsymbol{\beta}}(\sigma_u^2, \sigma^2) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}) \text{ and}$$

$$\hat{u}_d(\sigma_u^2, \sigma^2) = \mathbf{GZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

In EBLUP (empirical BLUP), the variances are estimated and plugged into the BLUP equation:

$$\hat{t}_{EBLUP} = \left(\sum_{k \in U_d} \mathbf{x}_k \right)' \hat{\boldsymbol{\beta}}(\hat{\sigma}_u^2, \hat{\sigma}^2) + N_d \hat{u}_d(\hat{\sigma}_u^2, \hat{\sigma}^2).$$

Another kind of EBLUP, here called EBLUP(Y) (Saei and Chambers, 2004), contains the conditional expectation of only that part of sum which is not observed in sample,

$$E\left(\sum_{k \in U_d - s_d} Y_k \,\middle|\, u \right) = \left(\sum_{k \in U_d - s_d} \mathbf{x}_k \right)' \boldsymbol{\beta} + (N_d - n_d) u_d.$$

The sample observations are included in the EBLUP(Y) estimator

$$\hat{t}_{EBLUP(Y)} = \left(\sum_{k \in U_d - s_d} \mathbf{x}_k \right)' \hat{\boldsymbol{\beta}}(\hat{\sigma}_u^2, \hat{\sigma}^2) + (N_d - n_d) \hat{u}_d(\hat{\sigma}_u^2, \hat{\sigma}^2) + \sum_{k \in s_d} y_k$$

EBLUP and EBLUP(Y) should have smaller MSE than GREG estimators, but they may have considerable design bias, especially if the design weights vary substantially.

The EBLUP estimators can be written using the predicted values

$$\hat{y}_k = \mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d$$

in forms resembling the synthetic estimator:

$$\hat{t}_{d;EBLUP} = \sum_{k \in U_d} \hat{y}_k(\hat{\sigma}_u^2, \hat{\sigma}^2)$$

For a logistic mixed model the EBP (empirical best predictor, e.g. Jiang and Lahiri, 2006) estimators are of the form

$$\hat{\bar{f}}_{d;EBP} = \sum_{k \in U_d - s_d} \hat{p}_k(\hat{\sigma}_u^2, \hat{\sigma}^2) \tag{15}$$

$$\hat{\bar{f}}_{d;EBP(Y)} = \sum_{k \in U_d - s_d} \hat{p}_k(\hat{\sigma}_u^2, \hat{\sigma}^2) + \sum_{k \in s_d} v_k\,, \tag{16}$$

where predictions are

$$\hat{p}_k = \exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d)/(1 + \exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d)\,.$$

3.4 Transformations of predictions

The synthetic estimator of a poverty indicator constructed from predictions is usually biased, in part due to the transformation of observations. As the income y is approximately distributed as lognormal, a model is fitted to $z_k = \log(y_k + 1)$, and the fitted values $\hat{z}_k$ are back-transformed to $\hat{y}_k = \exp(\hat{z}_k) - 1$. This should be followed by a bias correction. A RAST bias correction term $c_{RAST,d}$ (Ratio Adjusted by Sample

Total; Chambers and Dorfman, 2003, Fabrizi et al., 2007b) would be chosen in each domain *d* so that the weighted sample sum of $c_{RAST,d}\hat{y}_k$ over the domain equals the weighted domain sample sum of the original incomes y_k.

However, RAST correction merely corrects the mean of predictions without affecting significantly their spread. It ignores the fact that the tails of the distribution of incomes usually contribute significantly to a poverty indicator. For example, the quintile share incorporates the first and last quintiles. Unfortunately, the distribution of predictions is concentrated around the average and the income distribution derived from the predictions is unrealistically even. Therefore, synthetic estimates of Gini coefficient and poverty gap tend to be too small and quintile share estimated from predictions is often too large. Moreover, the differences between synthetic domain estimates are too small. We introduce linear and non-linear transformations as generalizations of the RAST correction.

We transform predictions so that they have similar histogram as the observed values. The transformation incorporates design weights even when they cannot be used in fitting the model, as is the case in many current R packages. This may reduce the design bias.

Consider predictions $\hat{y}_k$ for units in population domain *d* ($k \in U_d$). We compare the distributions of predictions and sample values by differences of percentiles. The percentiles of the $\hat{y}_k$ ($k \in U_d$) are denoted by $\hat{p}_{cd}$, $1 \le c \le 99$. The corresponding percentiles of the sample values y_k ($k \in s_d$), denoted by p_{cd}, are obtained from the HT estimate of the cumulative distribution function. Thus design weights contribute to the procedure. Our goal is to find a linear transformation defined by parameters a_d and b_d so that the percentiles of "expanded predictions" $y_k^* = a_d + b_d\hat{y}_k$ are close to corresponding percentiles p_{cd} of observations. Let p_{cd}^* denote the *c*th percentile of y_k^*, $k \in U_d$. We minimize the differences between the percentiles p_{cd}^* and p_{cd}:

$$S = \sum_c \left(p_{cd}^* - p_{cd} \right)^2 .$$

By noting that $p_{cd}^* = a_d + b_d \hat{p}_{cd}$ we obtain

$$S = \sum_c \left(a_d + b_d \hat{p}_{cd} - p_{cd} \right)^2 .$$

Obviously, S is minimized for parameters a_d and b_d by OLS corresponding to a linear regression model with $x = \hat{p}_{cd}$ and $y = p_{cd}$. The transformed domain predictions are

$$y_k^* = \hat{a} + \hat{b}\hat{y}_k . \qquad (17)$$

Weak auxiliary information may lead to negative transformed predictions (17). Here we outline a procedure for avoiding negative values. We derive non-linearly transformed predictions $\tilde{y}_k$ with percentiles of $\log(\tilde{y}_k)$, $k \in U_d$, close to corresponding percentiles of $\log(y_k)$, $k \in s_d$. As the percentiles of log-transformed vectors are logarithms of the original percentiles (although this does not always hold for the median), we minimize

$$\sum_c \left(a_d + b_d \log(\hat{p}_{cd}) - \log(p_{cd}) \right)^2 .$$

The parameters a_d and b_d are again found by OLS. Expanded predictions $\tilde{y}_k$ are then defined by

$$\log(\tilde{y}_k) = \hat{a}_d + \hat{b}_d \log(\hat{y}_k),$$

that is,

$$\tilde{y}_k = \exp\left(\hat{a}_d + \hat{b}_d \log(\hat{y}_k) \right). \qquad (18)$$

These expanded predictions are never negative. The log-transformation appears more natural for log-normally distributed observations than (17). For practical purposes, function $\log(x+1)$ was applied instead of $\log(x)$. However, the proportion of negative or zero incomes should not exceed 1%, to avoid undefined logarithms.

In a small domain, there is not enough data for reliable estimation of the percentiles of observations, and consequently the estimated parameters in the transformation (18) are inaccurate. With the Finnish data set we decided to calculate the p_{cd} from the whole sample instead of each domain, but such a procedure may result in bias. With the Amelia data, we obtained better results by minimizing the following sum over domains *d*:

$$\sum_d \sum_c \left(a_d + b \log\left(\hat{p}_{cd}\right) - \log\left(p_{cd}\right)\right)^2$$

This amounts to fitting a linear fixed-effects model with domain-specific intercepts a_d and common slope *b*. The expansion transformation is then

$$\tilde{y}_k = \exp\left(\hat{a}_d + \hat{b} \log\left(\hat{y}_k\right)\right).$$

In the Amelia data, about 1.5% of the people had zero equivalized income (variable EDI2), and negative incomes did not occur. In order to take the zeroes into account, we incorporated zero predictions into the transformation as follows. Let p_0 denote the proportional frequency of zero among the equivalized incomes in the sample. In a sorted vector of N_d domain predictions, roughly $N_d p_0$ smallest elements are replaced by zero. Then the percentiles $\hat{p}_{cd}$ are calculated from the positive predictions and the p_{cd} are calculated from positive sample values. Transformation (18) is applied only to the positive predictions, and zero predictions are included in the estimator.

To account for negative income values, we propose that the log-transformation in (18) is performed by function $\log(x+c+1)$, where c is the absolute value of the minimum over all observations and predictions, if negative observations or predictions occur and $c=0$ otherwise. Zero observations are then not treated separately, and all observations and predictions contribute to the percentiles. Instead of function $\exp$, we would apply $f(x)=\exp(x)-(c+1)$ in (18). This approach is adopted in the R algorithms, but it was not necessary in the simulation experiments, as negative incomes did not appear.

The range of percentage points may have large impact on the estimator. The percentiles are calculated at c=1, 2,…,99 for quintile share and Gini coefficient. For poverty gap, we used c=1, 2,…,50 in Table 12 and with Amelia data, but in tables 13-16 we used percentiles up to the poverty line. If the data are suspected of containing a lot of outliers, their effect is probably reduced by excluding some of the largest percentiles.

If the model incorporates few auxiliary variables, the number of distinct predictions is small, and the histogram of expanded predictions will consist of few bars, representing a poor approximation of the true distribution. When some of the auxiliary variables also define the domains, this problem is pronounced. For example, if the domains are defined by country, gender and age class, then with x-variables gender, age class and urbanisation, predictions in each domain have only three distinct values corresponding to the classes of urbanisation. Then the predictor involving expanded predictions may not yield good results.

3.5 Frequency-calibrated predictors calculated using known domain marginal totals of auxiliary variables

We develop here a new method that may be feasible in situations where only aggregate-level auxiliary data are available. Suppose that only the totals of auxiliary variables are known in a domain of population. In the case of qualitative x-variables, this means that the domain sizes and domain frequencies of classes are known in the

population; in other words, the totals of class indicators are known. From demographical population registers, we probably obtain at least the domain frequencies of classes for each combination of gender and age class. To calculate a predictor of a poverty indicator, we seemingly need the predictions for all population units i.e. access to the population data at the unit level. But actually, it is enough to know the frequencies of distinct values of predictions in a domain. We pursue this goal.

When a mixed model is fitted to log-transformed observations, the back-transformed predictions have the form

$$\hat{y}_k = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_{1k} + ... + \hat{\beta}_p x_{pk} + \hat{u}_d\right) - 1 = \exp\left(\hat{\beta}_0 + \hat{u}_d\right) \prod_i \exp\left(\hat{\beta}_i x_{ik}\right) - 1,$$

a nonlinear function of the values x_{ik}. Therefore we cannot derive the frequencies of distinct values of $\hat{y}_k$ from the known marginal totals. It appears necessary to have access to the frequencies of distinct values of $\mathbf{x}_k = \left(x_{1k}, x_{2k}, ..., x_{pk}\right)$ in each domain. We propose a method of estimating these frequencies using the design weights, the $\mathbf{x}_k$ in the sample and the known marginal totals.

Consider domain *d*. Denote the set of observed distinct values of $\mathbf{x}_k$, $k \in s_d$, by

$$X_d = \left\{z_1, z_2, ..., z_m\right\}.$$

A direct estimate of the domain frequency of $z \in X_d$ is

$$\hat{n}_z = \sum_{k \in s_d} a_k I_{x_k = z}.$$

These frequencies do not, in general, sum up to the known marginal totals. This requirement is formulated as a calibration equation

$$\sum_{k \in U_d} x_k = \sum_{z \in X_d} n_z z = t_d \ .$$

Calibration is used to obtain new frequencies $\hat{n}_z^*$ that are close to the $\hat{n}_z$ and also satisfy the calibration equations. As a measure of distance of $\hat{n}^* = (\hat{n}_z^*; z \in X_d)$ to $\hat{n} = (\hat{n}_z; z \in X_d)$ we have used the chi-squared distance

$$\sum_{z \in X_d} \frac{\left(\hat{n}_z - \hat{n}_z^*\right)^2}{\hat{n}_z}.$$

This distance is minimized subject to the calibration equations

$$\sum_{z \in X_d} \hat{n}_z^* z = t_d$$

by

$$\hat{n}_z^* = \hat{n}_z \left(1 + \lambda_d' z\right), \tag{19}$$

where the Lagrange multiplier λ_d is

$$\lambda_d = \left(t_d - \sum_{z \in X_d} \hat{n}_z z\right)\left(\sum_{z \in X_d} \hat{n}_z z z'\right)^{-1}$$

Unfortunately, some of the $\hat{n}_z^*$ can be negative. In our simulations, the average proportion of negative estimates was smaller than 2% with the Finnish data set but about 10% in the Amelia data set. We replaced negative estimates by zero. After this, the calibration equations do not necessarily hold.

Negative frequencies might be avoided by distance measure

$$\sum_{z \in X_d} \hat{n}_z^* \log\left(\hat{n}_z^* / \hat{n}_z\right) - \hat{n}_z^* + \hat{n}_z \,,$$

which is minimized under the calibration equations at

$$\hat{n}_z^* = \hat{n}_z \exp\left(\lambda' z\right).$$

This solution is found by a fixed point iteration algorithm (Singh and Mohl, 1996) involving repeated iteration of

$$\lambda_{i+1} = \lambda_i + \left(t_d - b\left(\lambda_i\right)\right) A^{-1}\left(\lambda_i\right);$$
$$A\left(\lambda\right) = \sum_z \hat{n}_z \exp\left(\lambda' z\right) zz'$$
$$b\left(\lambda\right) = \sum_z \hat{n}_z z \exp\left(\lambda' z\right)$$

However, this algorithm failed to converge too often, and was not applied in simulations.

To avoid singular matrices, we excluded from each *z* the indicators of classes that did not appear in the sample domain. Moreover, if two auxiliary variables had identical values in a domain, the latter variable was removed. Corresponding modifications were made in the vector t_d. If the algorithm still failed due to linear dependencies of auxiliary variables, for example, we used the initial estimates $\hat{n}_z$. This occurred rarely. The vector of predictions in the domain is finally obtained by repeating the fitted value associated with each $z \in X_d$ in the domain $\hat{n}_z^*$ times (after rounding), and expansion by (18) is applied. We call the resulting predictor a frequency-calibrated, or an n-calibrated predictor.

We have described the algorithm assuming that all the auxiliary variables are qualitative. It is possible to use the algorithm also when some of the x-variables are quantitative. However, it is probably necessary to transform a quantitative variable to have few distinct values.

The algorithm can be applied even when some of the auxiliary totals are not known in the population. We have replaced unknown population marginals by their GREG

estimates. As an example, suppose the population frequencies of age classes, gender and labour force status classes are known but a better fitting model includes also the socio-economic status of the head of the household, which is unknown in the population. We substitute GREG estimates of the frequencies of socio-economic status classes for corresponding marginal totals in the algorithm. The frequency-calibrated predictors have benefitted from the inclusion of a good auxiliary variable although its marginal totals are estimated. The GREG estimators were assisted by a multinomial logistic fixed effects model (R function multinom in package nnet).

3.6 Composite estimators

A *composite (COMP) estimator* is constructed from two estimators, one typically design unbiased ($\hat{\theta}_1$) and the other with small variance ($\hat{\theta}_2$). The composite is defined as a linear combination of the estimators:

$$\hat{\theta}_{COMP} = \lambda\hat{\theta}_1 + (1-\lambda)\hat{\theta}_2; 0 \le \lambda \le 1. \qquad (20)$$

This is expected to combine the best properties of its components. The composite estimator should have small design bias and smaller variance than the unbiased component, over some usually unknown range of λ. The MSE of the composite estimator is minimized by

$$\hat{\lambda} = \frac{MSE(\hat{\theta}_2)}{MSE(\hat{\theta}_2) + MSE(\hat{\theta}_1)}.$$

In the case of an unbiased $\hat{\theta}_1$, the $MSE(\hat{\theta}_1)$ can be replaced by variance $Var(\hat{\theta}_1)$. If $\hat{\theta}_1$ is not design unbiased, the equation is still applicable but the composite estimator may have significant design bias.

In domain estimation, separate $\hat{\lambda}$ are calculated for each domain *d*. As there is a lot of variability in the estimated $\hat{\lambda}_d$ values, Rao (2003, p. 59) recommends using the

average of $\hat{\lambda}_d$ over domains. We compared empirically some approaches to averaging the $\hat{\lambda}_d$ values over a subset of domains, such as domains in same region or domains with similar size, and chose to use average weights over domain size classes. The domain size classes were defined by expected sample size, the class boundaries being 50 and 100 elements in our simulations.

The default (direct) estimator of a poverty indicator is not necessarily nearly design unbiased in small domains as the indicator is a non-linear function of equivalized incomes. We still used the direct estimator as the design unbiased component $\hat{\theta}_1$ in the composite estimator (20). It is usually difficult to derive the theoretical variance of $\hat{\theta}_1$, and therefore jackknife has been used (Leiten and Traat, 2006). We applied bootstrap variance estimation: An artificial population is generated by cloning each unit with frequency equal to rounded design weight. Bootstrap samples are drawn with the original sampling design from the artificial population. The variance of the default estimator is then estimated by the sample variance of estimates in the bootstrap samples. If the direct estimator has significant design bias, as in the case of poverty gap, a bootstrap MSE might be used instead.

In small-area estimation, the second component of a composite estimator is often a synthetic estimator, which has small variance. The synthetic estimator of a poverty indicator is obtained by calculating the indicator's value from predictions derived under the specified model. As the MSE of the synthetic estimator is unknown, it has been suggested (Rao, 2003, p. 52; Fabrizi et al., 2007a) that the MSE is estimated by

$$\hat{MSE}(\hat{\theta}_2) = \left(\hat{\theta}_2 - \hat{\theta}_1\right)^2 - \hat{MSE}(\hat{\theta}_1), \qquad (21)$$

where $\hat{MSE}$ denotes the estimator of MSE. This is a somewhat crude method, and $\hat{MSE}(\hat{\theta}_2)$ can even be negative. In simulations we have replaced negative estimates by 0; then the composite estimator equals the synthetic estimator.

An alternative approach for estimating the mean squared errors of the synthetic estimators is based on parametric bootstrap with an algorithm similar to a

corresponding algorithm in Molina and Rao (2010). Our algorithm is the following:

(1) Fit a mixed model *m* to the sample *s* from population *U*. The parameter β is estimated by $\hat{\beta}$. The variance of random effects is estimated by $\hat{\sigma}_u^2$ and the variance of errors is estimated by $\hat{\sigma}_e^2$.

(2) Generate a bootstrap population U_i by simulating the *y* values for the original population. Firstly, the random effects u_d^* are simulated from $N(0,\hat{\sigma}_u^2)$ for each domain *d*. The *y*-values are generated from the model *m*: $y_k^* = x_k'\hat{\beta} + u_{d(k)}^* + \varepsilon_k^*$, where *d(k)* is the domain containing the unit *k* and ε_k^* is simulated from $N(0,\hat{\sigma}_e^2)$. The *y* variable is the only difference between U_i and the original population *U*. All the other variables are identical in U_i and *U*.

(3) Take a sample s_i from U_i, using the indices of the original sample $s \subset U$. Then the *x*-variables of s_i are identical with the *x*-variables of *s*. This means that the derived MSE can be regarded as conditional given the *x*-variables.

(4) Fit a mixed model to s_i with the structure of model *m* and calculate estimates $\hat{\theta}_{id}$ in the domains.

(5) Calculate the true values θ_{id} in the bootstrap population U_i and calculate the squared errors $\left(\hat{\theta}_{id} - \theta_{id}\right)^2$.

(6) Repeat steps 2-5 100 times and collect the squared errors $\left(\hat{\theta}_{id} - \theta_{id}\right)^2$, *i*=1,2,...,100. Calculate MSE estimates as

$$\frac{1}{100}\sum_{i=1}^{100}\left(\hat{\theta}_{id} - \theta_{id}\right)^2 \qquad (22)$$

3.7 Simulation-based methods

The *conditional expectation* $E(t \mid y_s)$ of a statistic t given observations $y_s = \{y_k; k \in s\}$ has an important optimality property: it minimizes, in general, the MSE among functions of y_s. In the case of poverty indicators, the conditional expectation is not necessarily tractable, but it can be approximated by simulation-based methods.

Molina and Rao (2010) have studied the estimation of poverty indicators by conditional expectations given $y_k (k \in s)$. Suppose an indicator can be written as a sum of functions $f(y_k)$. If the conditional expectations $E\left(f(y_k) \mid y_s; x_k\right)$ were known, the indicator would be estimated by

$$\hat{t} = \sum_{i \in U-s} E\left(f(y_i) \mid y_s; x_i\right) + \sum_{k \in s} f(y_k).$$

Molina and Rao (2010) approximate the conditional expectation $E\left(f(y_i) \mid y_s; x_i\right)$ by an average over simulations from the conditional distribution. As the income is approximately lognormally distributed, the y_k are transformed to $z_k = g(y_k) = \log(y_k + 1)$ and the simulations are based on the conditional distribution of z_i given $z_s = \{z_k; k \in s\}$:

$$E\left(f\left(y_i\right) \mid y_s; x_i\right) = E\left(f\left(g^{-1}(z_i)\right) \mid z_s; x_i\right) \approx \frac{1}{K} \sum_{t=1}^{K} f\left(g^{-1}(u_{it})\right); \qquad (23)$$

u_{it} follows normal distribution given z_s. The parameters of the conditional distribution are replaced by their estimates.

When a poverty indicator $t = f(y_1, ..., y_N) = f(y_r, y_s)$, $y_r = \{y_k; k \in U - s\}$, cannot be expressed as a sum, it is estimated by an average over $f(y_{rt}^*, y_s)$ (t=1,...,K), where the

y_{rt}^* are simulated from their conditional distribution given y_s and *x* (Molina and Rao, 2010). This approach is applicable, for example, in estimation of poverty gap, which incorporates the median income of poor people.

These simulation-based methods resemble imputation (e.g. Rubin, 1987, Schafer, 1997, Münnich and Wiegert (2001) Laaksonen, 2002), where missing values for nonrespondents are replaced by values generated with the help of a model. In fact, any imputation method could be used to impute all values in the unknown part of the population, although this is not common practice. In *conditional mean imputation*, the unknown values are replaced by conditional expectations: predictions from the model are substituted for $y_k, k \notin s$. We call such estimators "synthetic", although the term is usually reserved for the sum of predicted values. We also use term “predictor”. In *random imputation*, values are simulated from the distribution specified by the model. In imputation based on a regression model fitted to the $z_k (k \in s)$, values of $z_i (i \notin s)$ are simulated from normal distributions $N(\hat{\mu}_i, \hat{\sigma}^2)$. In other words, a random error term distributed as $N(0, \hat{\sigma}^2)$ is added to the prediction. Although it seems counterintuitive that adding random error could yield benefits over conditional mean imputation, the resulting estimator may have at least smaller bias. When a mixed model has been fitted, the values in domain *d* are simulated from $N(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d, \hat{\sigma}^2)$. In the case of random imputation, it would be interesting to generate independent simulated y_{rt}^*-vectors (*t* = 1,…,*K*) as in *multiple imputation* (e.g. Rubin, 1987; Schafer, 1997; Bjornstad, 2007) and calculate the average of indicator values $f(y_{rt}^*, y_s)$ over the simulations.

We have investigated the applicability of the method of Molina and Rao by simulation experiments for the Finnish register data. To save time, we let the number of simulations to depend on domain sample size n_d as $K_d = 2000 / n_d$. In small domains, this choice reduces the variance of the average over simulations.

4 Estimators for poverty indicators and results of Monte Carlo simulation experiments

4.1 Introduction

We introduce here the estimators of poverty indicators and present numerical results based on Monte Carlo simulation experiments. We use design-based simulation methods. Empirical data are based on statistical registers maintained by Statistics Finland and on Amelia population generated by Alfons et al. (2011b). We discuss poverty rate, poverty gap, Gini coefficient and quintile share. Empirical properties (design bias and accuracy) are evaluated.

4.2 Experimental design

Design bias and accuracy of estimators of the selected poverty indicators were examined by design-based simulation. We used two populations: a partially register-based Finnish population and the synthetic Amelia population (Alfons et al., 2011b).

4.2.1 Register-based population from Western Finland

The artificial Finnish population of one million persons was constructed from income data of seven NUTS3 -regions in Western Finland. The household properties, such as demographic composition and equivalized income were obtained from registers. The values of auxiliary variables of the household heads were obtained from a household survey. Some personal auxiliary variables, most notably education level, had to be imputed for other members of each household; nonetheless, the population was realistic enough for a simulation study. Unless otherwise specified in a table caption, the tables present results for this population.

In the simulations, $K = 1000$ samples of $n = 5000$ persons were drawn from the unit-level population. We used unequal probability sampling in addition to equal

probability sampling. The sampling design was SRSWOR or PPS. For PPS, an artificial size variable was generated as a function of a qualitative variable. Then the PPS is approximately identical with stratified sampling. PPS was defined so that people with low income appear in samples with larger probability than people with large income. Therefore low education levels and certain socio-economic classes were given the largest inclusion probabilities.

In PPS based on education level, the classes and relative inclusion probabilities are as follows (*p* is a constant depending on class frequencies):

Education class	**Inclusion probability**
0	5*p*
3	5*p*
4	4*p*
5	3*p*
6	2*p*
7	*p*
8	*p*

For PPS by socio-economic status (socstrat), inclusion probabilities were defined as follows:

Socio-economic class	**Inclusion probability**	**Mean income**
1	p/2	85069
2	p/3	68328
3	p/5	76491
4	*p*	58520
5	*p*	62448
6	*p*	56862

The mean equivalized income varied quite a lot but not linearly as a function of the size variable.

Our domains were 36 NUTS4 regions or 70 cells in the cross-classification of NUTS3 region, gender and age class (0-15, 16-24, 25-49, 50-64, and 65- years). These domains were classified by the expected sample size to size-classes with class boundaries at 50 and 100.

The following auxiliary variables were used:

Variable	**Label**	**Codes**
Age class	Age	0-15, 16-24, 25-49, 50-64, and 65- years
Gender	Gender	1 Males, 2 Females
House ownership	Indicator showing when the household owns the dwelling	0 (No), 1 (Yes)
Educ-thh	The number of household members having tertiary educational level	Count
Education	Education level of the household head	0 (Lowest) to 8 (Highest)
Empmohh	The total number of months of all household members being employed	Count
Socstrat	Socio-economic status of HH head	1 Wage and salary earners 2 Farmers 3 Other entrepreneurs 4 Pensioners 5 Other categories 6 Not specified
Lfs-code	Employment status of HH member	1 Employed 2 Unemployed 3 Not in workforce

We created indicators for each class of a qualitative variable. The most commonly used model had auxiliary variables age and gender with interactions, socstrat and lfs-code. The corresponding linear fixed-effects model fitted to logarithms of income in the population had coefficient of determination $R^2 = 0.101$. When auxiliary variables house ownership and educ-thh were added to the model, the R^2 increased to 0.164.

4.2.2 Amelia population

From the synthetic Amelia data set constructed using SILC data (Alfons et al., 2011b), we drew samples with SRSWOR (n = 2000) and PPS (n = 6000) based on a size variable with value 3 for education levels (ISCED) 0-3 and 2 for others. Forty regions (variable DIS) were classified by expected sample size with class boundaries at 45 and 55. Demographic domains were defined by age, gender and NUTS2 regions. For poverty rate, these domains were classified by size with breakpoints 50 and 100, for poverty gap with breakpoints 20 and 30. Our models fitted to the logarithm of the equivalized income variable EDI2 incorporated age class and gender with interactions, attained education level (ISCED), activity (working, unemployed, retired, or otherwise inactive) and degree of urbanisation (three classes).

4.2.3 Quality measures

From each simulation consisting of K=1000 samples, the following quality measures, among others, were calculated for each domain estimator: mean, bias

$$Bias = \frac{1}{K}\sum_{k=1}^{K}(\hat{\theta}_{dk} - \theta_d), \tag{36}$$

absolute relative bias

$$ARB = \frac{\left|\frac{1}{K}\sum_{k=1}^{K}(\hat{\theta}_{dk} - \theta_d)\right|}{\theta_d} \tag{37}$$

and relative root mean squared error

$$RRMSE = \frac{\sqrt{\frac{1}{K}\sum_{k=1}^{K}(\hat{\theta}_{dk} - \theta_d)^2}}{\theta_d}. \qquad (38)$$

We present the averages of the quality measures over domain classes defined by domain size.

4.2.4 Contamination schemes

Outlier and contamination experiments were carried out as proposed in Hulliger and Schoch (2010), p. 7. In contamination experiments, outliers were created in each sample without modifying the population. Motivation for this choice is discussed in Alfons et al. (2011a). In OCAR (outlying completely at random), one percent of sampled persons were declared as outliers, chosen completely at random. In OAR (outlying at random), the probability of being an outlier varied as a function of labour force status and pensioner status (socstrat) as follows: 0.04 for employed people, 0.02 for the unemployed, 0.03 for people not in workforce but 0.01 for pensioners. In the Finnish population, the equivalized income of the outlier's household was the target of contamination, whereas in Amelia, the personal cash or near-cash income of an outlier was contaminated. Under CCAR contamination (contaminated completely at random), a normally distributed value from $N(500000, 10000^2)$ was added to the target income variable. Under NCAR (not contaminated at random), the outlier's income value was multiplied by 1000. Under OAR, the expectations of contamination $N(\mu, 10000^2)$ were 5,000,000 for the employed, 4000 for the unemployed, 90000 for people not in workforce but 200 for pensioners. In Amelia, the equivalized income in the outlier's household was calculated anew using other personal components and household-level components of the disposable income of the household. OAR contamination may sometimes result in negative incomes. In simulations these were unfortunately left out from model fitting, as R replaces logarithms of negative values by missing values.

4.2.5 Estimators

Most of the mixed models were fitted by R package **nlme** using maximum likelihood. Design weights were then not used. For Tables 4 and 5, we incorporated design weights into model fitting by glmer function of R package lme4. The lme4 package fits mixed models by a penalized, iteratively reweighted, least squares algorithm (Bates, 2011). The linear and logistic fixed-effects models were fitted with GWLS and maximum pseudolikelihood methods incorporating design weights.

In experiments with Gini coefficient, poverty gap and quintile share we compare the following estimators:

Estimator	**Description**	**Reference equations**
Default	The default (direct) estimator of the poverty indicator	(27), (30), (33)
Model-based estimators		
Predictor	Estimator calculated from predicted values	(28), (31), (34)
Expanded predictor	Estimator (28), (31), or (34) from transformed predictions; used equation in parentheses	(17) or (18)
n-calibrated predictor	Predictor type estimator based on calibrated frequencies of fitted values	(18) and (19)
Composite estimators		
Composite	Composite estimator incorporating default estimator and expanded predictor	(17) or (18), (20)
n-calibrated composite	Composite estimator incorporating default estimator and frequency-calibrated predictor	(18), (19), (20)

In the n-calibrated estimator, we treated socstrat as a variable unknown in the population. The marginal frequencies of socstrat classes were imputed by GREG assisted by multinomial logistic model (R module nnet). A more technical summary of the methods is given in Annex 2 and 3.

4.3 At-risk-of poverty rate

At-risk-of-poverty threshold is 60 % of the median equivalized income of persons in the whole population or in a regional population. In experiments, we used the population median. People whose income is below or at the at-risk-of-poverty threshold are referred to as “poor”. *At-risk-of-poverty rate* is the proportion of poor people in a domain.

To estimate the *reference median income* M, we first derive the HT estimator of the distribution function of equivalized income in the whole population. The distribution function of *y* in *U* is

$$F_U(t) = \frac{1}{N}\sum_{k\in U} I\{y_k \le t\}$$

This is estimated by HT:

$$\hat{F}_U(t) = \frac{1}{\hat{N}}\sum_{k\in s} a_k I\{y_k \le t\},$$

where the estimated population size is $\hat{N} = \sum_{k\in s} a_k$.

$\hat{M}$ is obtained from $\hat{F}_U$ as the smallest $y_k\,(k \in s)$ for which $\hat{F}_U(y_k) > 0.5$. In the special case of $\hat{F}_U(y_{(k)}) = 0.5$ for *k*th observation in sorted *y*, the median is the average of $y_{(k)}$ and $y_{(k+1)}$.

In poverty rate estimation, our goal is to estimate

$$R_d = \frac{1}{N_d}\sum_{k\in U_d} I\{y_k \le 0.6M\}.$$

4.3.1 HT-CDF estimator

HT-CDF estimator of poverty rate is based on the HT estimator of the distribution function. The distribution function is defined in domain U_d as

$$F_d(t) = \frac{1}{N_d} \sum_{k \in U_d} I\{y_k \le t\}.$$

It is estimated by

$$\hat{F}_d(t) = \frac{1}{\hat{N}_d} \sum_{k \in s_d} a_k I\{y_k \le t\},$$

where $\hat{N}_d = \sum_{k \in s_d} a_k$.

The poverty rate is then estimated by

$$\hat{r}_{d;HT} = \hat{F}_d(0.6\hat{M}). \qquad (24)$$

Problems arise if empty domains ($n_d = 0$) are common. Then $\hat{F}_d(t)$ might be replaced by an average of $\hat{F}_p(t)$ over domains p in neighbouring regions, but this would probably reduce differences between regions too much.

4.3.2 Methods based on poverty indicators

Poverty indicator $v_k = I\{y_k \le 0.6\hat{M}\}$ equals 1 for persons with income smaller than the estimated at-risk-of-poverty threshold and 0 for others. If $\hat{M}$ equals the true median income, v_k identifies the poor people. The poverty indicator is used in methods such as logistic GREG, model calibration, and EBP.

The HT estimate of the number of poor people is

$$\hat{f}_{d;HT} = \sum_{k \in s_d} a_k v_k$$

and the share of persons at-risk-of poverty is estimated by

$$\hat{r}_{d;HT} = \frac{\hat{f}_{d;HT}}{\hat{N}_d} \qquad (25)$$

or

$$\hat{r}_{d;HT} = \frac{\hat{f}_{d;HT}}{N_d}.$$

The form (25) is actually identical with HT-CDF. Corresponding LGREG estimators are

$$\hat{r}_{d;LGREG} = \frac{\hat{f}_{d;LGREG}}{\hat{N}_d}$$

and

$$\hat{r}_{d;LGREG} = \frac{\hat{f}_{d;LGREG}}{N_d}. \qquad (26)$$

In the simulation experiments, the denominator was N_d, and estimators (24) and (26) were used.

4.3.3 Simulation results

In our Monte Carlo simulation experiments, we compared the following estimators:

Estimator	Description	Equations
Default	The default (direct) estimator of the poverty rate	(24)
Design-based estimators		
GREG	Generalized regression estimator assisted by a linear fixed-effects model	(26)
LGREG	Logistic GREG estimator assisted by a logistic fixed-effects model	(26)
MLGREG	GREG estimator (26) assisted by a logistic mixed model	(26)
MC	Model calibration; equation in parentheses e.g.MC(10)	(10), (12), (13)
Model-based estimators		
LSYN	Synthetic estimator based on a logistic fixed effects model	(14)
EBP	Empirical Best Predictor type estimator based on a logistic mixed model	(15)
EBP(Y)	Alternative EBP type estimator based on a logistic mixed model	(16)

Table 2 compares poverty rate estimators assisted by fixed effects models. Section a) shows results for a common model formulation where the model does not account for domain differences. NUTS3 indicators are included in Section b) to account for regional variation. Section c) includes domain-specific fixed effects. In this case the model-based LSYN and model-assisted LGREG coincide. Under SRSWOR, it was not necessary to include design weights in model fitting.

Table 2. Poverty rate estimators assisted by logistic and linear fixed effects models.
Design: SRSWOR.
Qualitative x: house ownership, age class, gender, lfs-code.
Domains: NUTS3 by age by gender (D = 70 domains)

Estimator	BIAS			ARB (%)			RRMSE (%)		
	minor	medium	major	minor	medium	major	minor	medium	major
a) Common model formulation									
Default	–0.04	0.03	–0.06	1.23	0.94	0.71	51.83	32.0	22.29
LSYN	–1.11	–0.32	0.54	13.95	12.39	5.87	18.16	16.54	10.15
LGREG	0.03	0.03	–0.05	0.68	0.87	0.65	48.55	30.55	20.66
GREG	0.03	0.03	–0.05	0.76	0.88	0.65	48.89	30.86	20.89
MC(12)	0.03	0.03	–0.05	0.68	0.86	0.65	48.56	30.55	20.65
MC(13)	0.03	0.03	–0.05	0.75	0.85	0.67	48.39	30.51	20.63
MC(10)	–0.12	0.01	–0.06	1.73	0.89	0.68	52.94	31.3	20.88
b) NUTS3 indicators added to x-variables									
Default	–0.06	0.02	–0.07	1.21	0.93	0.73	51.82	31.98	22.29
LSYN	–0.01	0.12	–0.14	7.98	8.05	4.9	19.35	15.99	11.41
LGREG	0.02	0.02	–0.06	0.71	0.84	0.68	48.4	30.5	20.66
GREG	0.02	0.02	–0.06	0.79	0.86	0.67	48.74	30.81	20.88
MC(12)	0.02	0.02	–0.06	0.72	0.84	0.68	48.42	30.51	20.64
MC(13)	0.02	0.02	–0.06	0.73	0.83	0.69	48.39	30.5	20.64
MC(10)	–0.15	0.0	–0.07	1.83	0.89	0.7	52.85	31.29	20.88
c) Domain indicators added to x-variables									
Default				1.21	0.93	0.73	51.82	31.98	22.29
LSYN				1.18	0.83	0.7	50.98	30.9	20.81
LGREG				1.18	0.83	0.7	50.98	30.9	20.81
GREG				1.08	0.85	0.67	50.84	31.1	20.98
MC(12)				1.15	0.82	0.7	51.04	30.93	20.81
MC(13)				1.09	0.84	0.7	50.7	30.95	20.83
MC(10)				1.89	0.88	0.7	52.45	31.26	20.88

The default estimator, model calibration (MC), and GREG estimators are nearly design unbiased. Among these methods, model calibration based on (13) has the smallest RRMSE. In (13), the sums of fitted values were calibrated at NUTS3 level. Therefore there is not much difference between models (a) and (b). LSYN had the smallest RRMSE but it was design biased.

A logistic mixed model is used next to compare model-based EBP with model-assisted MLGREG (Table 3). Domain differences are accounted for by regional-level (Section a) or domain-specific (Section b) random intercepts in the model. In both cases, the EBP estimator has large negative design bias, especially for small domains, and MLGREG appears nearly design unbiased as expected. However, EBP shows better accuracy than MLGREG and other nearly unbiased methods of Table 2. MLGREG has somewhat larger bias than LGREG.

Table 3. Poverty rate estimators assisted by a logistic mixed model.
Design: SRSWOR.
Qualitative x: house ownership, lfs-code, age class, gender.
Domains: NUTS3 by age by gender.
Mixed model with NUTS3 random intercepts was fitted by nlme.

Estimator	BIAS			ARB (%)			RRMSE (%)		
	minor	medium	major	minor	medium	major	minor	medium	major
a) NUTS 3 level random intercepts									
EBP(Y)	–1.47	–0.53	0.02	14.85	10.75	4.07	20.83	17.22	10.81
MLGREG	0.01	0.03	–0.05	0.66	0.87	0.68	48.66	30.72	20.75
b) Domain-specific random intercepts									
EBP(Y)	–1.43	–0.55	0.16	14.75	8.96	3.99	22.49	19.26	14.54
MLGREG	0.28	0.13	–0.27	2.2	3.44	2.76	55.67	39.87	30.44

From tables 2 and 3 we see that random intercepts or fixed effects associated with NUTS3 regions yield better results than domain-specific effects.

Tables 4 and 5 show the effect of incorporating the design weights in fitting a mixed model. If the variable socstrat determining the size variable in PPS is not included in the model (Table 4), using design weights in fitting (EBP(Y)-W, no socstrat) results in smaller bias and RRMSE than model fitting without weights (EBP(Y), no socstrat). When socstrat was included in the model, EBP(Y)-W had smaller design bias than EBP(Y) but slightly larger RRMSE. MLGREG did not yield as small RRMSE as EBP(Y), but it had smaller bias. MLGREG-W benefitted slightly from using design weights in model fitting. We draw similar conclusions from Table 5.

Table 4. Poverty rate estimators with design weights incorporated in model fitting (lme4) in methods with suffix "W".
Design: PPS based on socstrat.
Qualitative x: age and gender with interactions, lfs-code and socstrat.
Domains: NUTS3 by age by gender
Logistic mixed model with NUTS3 random intercepts was fitted by lme4.

Estimator	ARB (%)				RRMSE (%)			
	minor	medium	major	all	minor	medium	major	all
Default	1.60	1.13	0.54	1.17	54.18	30.21	20.95	36.79
EBP(Y)	11.84	8.21	5.01	8.82	19.73	15.61	11.63	16.23
EBP(Y), no socstrat	13.40	9.88	7.37	10.60	20.94	16.94	12.93	17.51
EBP(Y)-W	9.33	8.04	5.57	7.97	20.00	16.23	12.39	16.76
EBP(Y)-W, no socstrat	9.58	8.27	5.47	8.14	20.01	16.38	12.43	16.83
MLGREG	1.56	1.13	0.59	1.17	53.95	30.22	20.89	36.69
MLGREG-W	1.57	1.14	0.58	1.17	53.64	30.12	20.82	36.53

Table 5. Poverty rate estimators in Amelia. Design weights are incorporated in model fitting (lme4) in methods with suffix "W".
Design: PPS based on ISCED.
Qualitative x: age and gender with interactions, ISCED, activity, and degree of urbanisation.
Domains: NUTS2 by age by gender.
Logistic mixed model with NUTS2 random intercepts was fitted by lme4.

Estimator	ARB (%)				RRMSE (%)			
	minor	medium	major	all	minor	medium	major	all
Default	0.76	0.61	0.32	0.67	29.14	23.08	17.36	26.09
EBP(Y)	8.29	9.25	7.78	8.56	13.50	13.77	10.92	13.36
EBP(Y), no ISCED	8.67	10.01	7.88	9.04	13.98	14.52	11.51	13.93
EBP(Y)-W	8.35	8.93	7.92	8.50	13.61	13.65	11.13	13.40
EBP(Y)-W, no ISCED	8.30	8.96	7.77	8.47	13.68	13.69	11.02	13.44
MLGREG	0.74	0.57	0.29	0.64	28.13	22.34	16.90	25.21
MLGREG-W	0.74	0.57	0.29	0.64	28.12	22.34	16.89	25.21

Table 6 shows how contamination affects poverty rate estimators. A robust method of fitting the logistic mixed model was not available. Nevertheless, the poverty rate estimators are fairly robust. Only when the proportion of outliers is 15%, bias especially is large. EBP(Y) has the smallest RRMSE in this experiment. It was also least affected by contamination.

Table 6. Poverty rate in contaminated data.
Design: SRSWOR.
Qualitative x-variables: age and gender with interactions, lfs-code and socstrat.
Domains: NUTS3 by gender and age class (70 domains).
Logistic mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme.

Estimator	ARB (%)				RRMSE (%)			
	minor	medium	major	all	minor	medium	major	all
Baseline (no contamination)								
Default	1.11	1.04	0.50	0.94	51.94	31.79	22.04	36.76
MLGREG	1.41	0.98	0.42	1.01	48.83	30.81	20.93	34.99
EBP(Y)	9.00	8.21	5.36	7.84	19.91	17.51	12.51	17.23
OCAR-CCAR 1%								
Default	1.69	1.29	0.50	1.25	52.13	31.93	22.11	36.90
MLGREG	1.91	1.25	0.45	1.30	49.04	30.94	21.00	35.13
EBP(Y)	8.47	8.52	5.33	7.77	19.68	17.73	12.54	17.24
OCAR-NCAR 1%								
Default	1.65	1.31	0.50	1.25	52.16	31.91	22.11	36.90
MLGREG	1.94	1.27	0.44	1.32	49.09	30.94	21.00	35.15
EBP(Y)	8.48	8.53	5.34	7.78	19.77	17.76	12.54	17.28
OAR-CAR								
Default	1.73	1.10	0.63	1.22	52.17	31.93	22.09	36.91
MLGREG	1.88	1.10	0.58	1.26	49.06	30.93	20.95	35.12
EBP(Y)	8.66	8.58	5.45	7.89	19.67	17.76	12.60	17.26
OCAR-CCAR 15%								
Default	23.36	15.81	4.93	16.02	63.25	39.23	23.45	44.20
MLGREG	23.72	15.84	4.92	16.16	60.43	38.25	22.46	42.56
EBP(Y)	21.24	20.04	6.17	17.30	28.87	27.20	13.72	24.71

Table 7 shows how contamination affects estimators under PPS. The bias of EBP(Y) is larger than in Table 6, with the exception of contamination of 15%. The RRMSE of other methods are larger than under SRSWOR.

Table 7. Poverty rate in contaminated data.
Design: PPS by socio-economic status.
Qualitative x-variables: age and gender with interactions, lfs-code and socstrat.
Domains: NUTS3 by gender and age class (70 domains).
Logistic mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme.

Estimator	ARB (%)				RRMSE (%)			
	minor	medium	major	all	minor	medium	major	all
Baseline (no contamination)								
Default	1.60	1.13	0.54	1.17	54.18	30.21	20.95	36.79
MLGREG	1.53	1.15	0.57	1.16	53.99	30.20	20.83	36.69
EBP(Y)	11.85	8.76	6.97	9.48	20.26	16.20	12.76	16.91
OCAR-CCAR 1%								
Default	2.04	1.39	0.63	1.46	54.33	30.28	20.94	36.87
MLGREG	2.13	1.41	0.64	1.50	54.14	30.26	20.83	36.77
EBP(Y)	11.41	8.68	7.24	9.35	20.01	16.15	12.92	16.84
OCAR-NCAR 1%								
Default	2.01	1.37	0.68	1.45	54.38	30.29	20.97	36.90
MLGREG	2.10	1.39	0.70	1.50	54.17	30.29	20.85	36.79
EBP(Y)	11.41	8.71	7.26	9.36	20.04	16.21	12.96	16.88
OAR-CAR								
Default	2.35	1.20	0.86	1.54	54.29	30.14	20.94	36.79
MLGREG	2.33	1.23	0.81	1.53	54.04	30.13	20.82	36.67
EBP(Y)	11.49	8.81	7.61	9.51	20.00	16.25	13.18	16.93
OCAR-CCAR 15%								
Default	21.53	14.86	10.08	16.22	63.75	36.99	26.08	44.21
MLGREG	21.97	14.78	10.16	16.36	63.66	36.94	25.99	44.14
EBP(Y)	17.73	16.87	12.57	16.26	25.95	23.20	18.57	23.19

Logistic mixed models are at least theoretically preferable to fixed effects models as they describe differences between domains parsimoniously. Model calibration (13) had small design bias and RRMSE with fixed effects models. Of all the poverty rate estimators, EBP might be the best choice unless it is important to avoid design bias. Our findings are similar to the conclusions of Fabrizi et al. (2007a) and Judkins and Liu (2000).

4.4 The Gini coefficient

Consider a population domain U_d of size N_d where the equivalized incomes are ordered: $y_{(1)} \le y_{(2)} \le ... \le y_{(N_d)}$.

The *Lorenz curve* $L_d(.)$ in domain *d* is defined at points k/N_d for persons $k \in U_d$ by

$$L_d\left(\frac{k}{N_d}\right) = \frac{\sum_{i \le k; i \in U_d} y_{(i)}}{\sum_{t \in U_d} y_t}.$$

The *x*-coordinate represents the first *k* persons' numerical proportion of the population and *y*-coordinate represents their proportion of the total income. For practical purposes, we define the Lorenz curve as a piecewise linear function, approximated by a line between consecutive points for k/N_d and $(k+1)/N_d$. If the income were uniformly distributed, the curve would be a line from (0,0) to (1,1). In real data, the Lorenz curve is below this line.

The *Gini coefficient* G_d in domain *d* is defined as

$$G_d = 1 - 2\int_0^1 L_d(x)dx.$$

With uniform income distribution, $G_d = 0$. Typical values for a country range from 0.2 to 0.4.

For a sample domain s_d, an HT-based estimate of the Lorenz curve is defined by first ordering the persons in the sample by equivalized income, $y_{(1)} \le y_{(2)} \le ... \le y_{(n_d)}$. The weights are correspondingly ordered by the income; the design weight of the observation at *i*th position in the ordered sample is denoted by a_i^s. Consider *k* first

persons in the ordered sample. Their numerical proportion of the population is estimated by

$$\frac{\sum_{i\le k;i\in s_d} a_i^s}{\sum_{t\in s_d} a_t}.$$

The proportion of the first *k* incomes of the total income is estimated by a ratio of two HT estimates:

$$\frac{\sum_{i\le k;i\in s_d} a_i^s y_{(i)}}{\sum_{t\in s_d} a_t y_t}$$

Thus, at a point for *k*, the Lorenz curve's HT-type estimator is defined by

$$L_{HT;d}\left(\frac{\sum_{i\le k;i\in s_d} a_i^s}{\sum_{t\in s_d} a_t}\right)=\frac{\sum_{i\le k;i\in s_d} a_i^s y_{(i)}}{\sum_{t\in s_d} a_t y_t}.$$

For integration, consecutive points are joined by lines. We have numerically verified that the *default* (direct) *estimator* $G_{HT;d}$ of the Gini coefficient for domain *d* is then equivalent to

$$G_{HT;d}=1-2\int_0^1 L_{HT;d}(x)dx. \qquad (27)$$

For domains with a single observation, the estimates are obtained from the whole country instead. Another viable option might be the synthetic estimator discussed next.

The synthetic estimator of the Lorenz curve is calculated using the ordered predicted incomes in population, $\hat{y}_{(1)} \le \hat{y}_{(2)} \le ... \le \hat{y}_{(N_d)}$:

$$L_{SYN;d}\left(\frac{k}{N_d}\right)=\frac{\sum_{i\le k;i\in U_d}\hat{y}_{(i)}}{\sum_{t\in U_d}\hat{y}_t}.$$

The *synthetic estimator* $G_{SYN;d}$ of the Gini coefficient for domain *d* is

$$G_{SYN;d}=1-2\int_0^1 L_{SYN;d}(x)dx \qquad (28)$$

We tried composite estimation of the Lorenz curve by a linear combination of $L_{HT;d}$ and $L_{SYN;d}$ of type (20) but it did not yield as good results as composite estimators incorporating $G_{HT;d}$ and $G_{SYN;d}$.

Table 8 shows an experimental comparison of the expanded predictor (17) of the Gini coefficient, the default estimator and the ordinary predictor (28). Benefits from the expansion (17) are obvious.

Table 8. Estimators of Gini coefficient assisted by linear mixed model.
Design: SRSWOR.
Quantitative x: educ-thh, empmohh.
Qualitative x: house ownership, lfs-code, socstrat.
Domains: 36 NUTS4 regions.
Mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme.

Estimator	BIAS			ARB (%)			RRMSE (%)		
	minor	medium	major	minor	medium	major	minor	medium	major
Default	–.007	–0.004	–.002	2.92	1.57	0.66	14.09	11.42	7.66
Predictor	–.066	–0.066	–.063	27.96	28.14	26.18	28.12	28.30	26.34
Expanded predictor (17)	–.004	–0.003	–.005	3.97	3.04	3.44	4.43	3.56	3.86
Composite	–.005	0.001	–.004	3.46	2.11	2.56	5.79	4.29	3.91

Tables 9 and 10 summarize experiments with contamination. The expanded predictor and frequency-calibrated predictor are better methods than the default one. They are

also fairly robust. Composite estimators have large design bias in the most contaminated data. In OCAR-NCAR, the bias and RRMSE of expanded predictor and frequency-calibrated estimator are larger under PPS than under SRSWOR.

Table 9. Gini coefficient in contaminated data.
Design: SRSWOR.
Qualitative x-variables: age and gender with interactions, lfs-code and socstrat.
Domains: NUTS4.
Mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme.

Estimator	ARB (%)				RRMSE (%)			
	minor	medium	major	all	minor	medium	major	all
No contamination								
Default	3.27	1.74	0.66	1.56	14.28	11.36	7.57	10.40
Expanded predictor (18)	4.55	6.37	3.12	4.94	5.10	6.68	3.72	5.39
Composite	2.27	3.59	2.18	2.90	6.28	5.94	3.70	5.18
Predictor	49.38	50.15	48.74	49.53	49.72	50.49	49.05	49.86
n-calibrated predictor	3.06	4.64	2.95	3.81	5.06	5.64	3.70	4.86
n-calibrated composite	2.46	2.84	2.09	2.51	6.09	5.37	3.64	4.85
OCAR-CCAR 1%								
Default	14.76	17.67	17.64	17.26	33.52	29.92	22.66	27.80
Expanded predictor (18)	13.15	14.98	9.38	12.70	13.34	15.14	9.63	12.90
Composite	13.00	15.66	12.00	13.97	18.59	18.36	13.10	16.49
Predictor	49.92	50.68	49.29	50.07	50.07	50.83	49.44	50.22
n-calibrated predictor	8.93	12.59	8.83	10.73	10.23	13.15	9.20	11.32
n-calibrated composite	10.50	14.11	11.66	12.72	16.53	16.95	12.84	15.41
OCAR-NCAR 1%								
Default	98.84	151	231	173	173	212	254	223
Expanded predictor (18)	15.31	17.08	11.26	14.73	15.73	17.49	11.77	15.18
Composite	68.85	111	193	135	116	153	212	169
Predictor	48.12	48.85	47.64	48.31	48.98	49.70	48.43	49.14
n-calibrated predictor	10.78	14.53	10.67	12.61	12.27	15.34	11.30	13.45
n-calibrated composite	63.22	107	192	131	108	148	211	165
OAR-CAR								
Default	88.07	118	141	122	139	152	151	150.05
Expanded predictor (18)	25.32	27.29	19.72	24.28	25.48	27.45	19.90	24.45
Composite	68.17	91.56	113	96.21	102.31	114	122	115
Predictor	32.54	33.10	31.96	32.61	34.44	34.99	33.66	34.43
n-calibrated predictor	20.67	24.79	19.07	22.15	21.59	25.25	19.35	22.61
n-calibrated composite	64.93	90.04	113	95.00	97.92	113	122	114

Table 10. Gini coefficient in contaminated data under PPS.
Design: PPS by socio-economic status.
Qualitative x-variables: age and gender with interactions, lfs-code and socstrat.
Domains: NUTS4.
Mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme without design weights.

Estimator	ARB (%)				RRMSE (%)			
	minor	medium	major	all	minor	medium	major	all
No contamination								
Default	4.11	2.40	0.84	2.08	16.58	13.17	8.58	11.99
Expanded predictor (18)	4.55	6.33	3.11	4.92	5.11	6.65	3.74	5.38
Composite	1.88	3.19	2.15	2.63	7.55	6.33	3.96	5.64
Predictor	47.58	48.22	47.04	47.70	47.65	48.29	47.11	47.77
n-calibrated predictor	3.32	4.75	3.00	3.92	5.05	5.52	3.72	4.80
n-calibrated composite	2.72	2.61	2.09	2.44	7.27	5.76	3.91	5.30
OCAR-CCAR 1 %								
Default	13.37	17.00	17.37	16.63	33.59	30.93	23.34	28.56
Expanded predictor (18)	12.48	14.18	8.76	11.99	12.73	14.42	9.11	12.26
Composite	11.87	14.90	11.35	13.20	18.13	17.88	12.65	16.03
Predictor	47.96	48.59	47.44	48.09	48.04	48.66	47.52	48.16
n-calibrated predictor	7.90	12.01	8.48	10.17	9.38	12.50	8.91	10.77
n-calibrated composite	9.21	13.44	11.17	12.03	16.06	16.57	12.52	15.04
OCAR-NCAR 1 %								
Default	93.11	149.85	229.28	170.65	168.79	211.05	251.31	219.72
Expanded predictor (18)	21.90	23.84	17.22	21.18	24.74	26.66	20.09	24.02
Composite	69.35	113.65	194.25	136.60	120.81	157.17	212.86	172.23
Predictor	46.54	47.11	46.19	46.70	46.68	47.25	46.33	46.84
n-calibrated predictor	16.26	21.06	16.87	18.88	20.06	24.17	19.86	22.04
n-calibrated composite	64.69	110.71	193.96	134.38	114.04	153.48	212.57	169.34
OAR-CAR								
Default	69.91	100.96	132.73	108.12	127.63	144.19	148.67	143.51
Expanded predictor (18)	24.46	26.30	19.01	23.41	24.63	26.47	19.20	23.59
Composite	55.00	77.76	102.93	83.69	92.71	105.97	114.76	107.30
Predictor	30.66	31.06	30.22	30.70	30.78	31.18	30.34	30.82
n-calibrated predictor	19.27	23.86	18.62	21.33	20.29	24.28	18.89	21.78
n-calibrated composite	51.80	76.02	102.87	82.35	87.90	103.80	114.71	105.53

Table 11 shows results with the Amelia dataset. Conclusions are similar as above.

Table 11. Gini coefficient in contaminated Amelia data under SRSWOR.
Qualitative x-variables: age and gender with interactions, ISCED, activity and degree of urbanisation.
Domains: districts (DIS).
Mixed model with DIS random intercepts was fitted to log(income+1) by lme without design weights.

Estimator	ARB (%)				RRMSE (%)			
	minor	medium	major	all	minor	medium	major	all
No contamination								
Default	2.60	2.10	1.61	2.11	12.89	11.64	10.43	11.68
Expanded predictor (18)	10.69	8.37	7.71	8.89	11.51	9.33	8.69	9.81
Composite	5.46	4.21	4.01	4.53	8.05	6.76	6.14	6.98
Predictor	21.75	23.27	23.74	22.94	22.59	24.00	24.44	23.70
n-calibrated predictor	6.22	4.31	3.74	4.72	11.32	9.06	7.84	9.40
n-calibrated composite	3.70	2.33	2.03	2.66	8.58	7.10	6.15	7.28
OCAR-CCAR 1 %								
Default	7.77	8.75	9.79	8.74	21.48	20.63	19.94	20.70
Expanded predictor (18)	12.65	10.29	9.63	10.81	13.38	11.12	10.50	11.63
Composite	10.82	9.59	9.51	9.94	13.32	11.91	11.46	12.21
Predictor	22.25	23.75	24.22	23.43	23.09	24.49	24.92	24.18
n-calibrated predictor	7.80	6.00	5.53	6.41	12.46	10.09	8.96	10.49
n-calibrated composite	7.66	6.71	6.70	6.99	12.13	10.67	9.99	10.92

4.5 Poverty gap

Relative median at-risk-of poverty gap, or *poverty gap* for short, in a region describes the difference between the poor people's median income and the at-risk-of-poverty threshold *t*. The threshold is usually estimated for the whole country. The poverty gap g_d in domain *d* is defined as a ratio

$$g_d = \frac{t - Md\{y_k; y_k \le t; k \in U_d\}}{t}. \qquad (29)$$

The *default* (direct) *estimator* $\hat{g}_d$ for domain *d* is calculated from the sample:

$$\hat{g}_d = \frac{\hat{t} - Md\{y_k; y_k \le \hat{t}; k \in s_d\}}{\hat{t}}. \qquad (30)$$

The *synthetic estimator* $\hat{g}_{d;SYN}$ of the poverty gap for domain *d* is calculated from the predicted values $\hat{y}_k$ after classifying people as poor when their predictions are below the estimated threshold $\hat{t}$:

$$\hat{g}_{d;SYN} = \frac{\hat{t} - Md\{\hat{y}_k; \hat{y}_k \le \hat{t}; k \in U_d\}}{\hat{t}} \qquad (31)$$

As the predictions vary less than the true incomes, the synthetic poverty gap estimate is usually too small.

Composite estimator (20) of the poverty gap incorporates the default estimator and the synthetic estimator:

$$\hat{g}_{d;COMP} = \hat{\lambda}_d \hat{g}_d + (1 - \hat{\lambda}_d)\hat{g}_{d,SYN}, \qquad (32)$$

where $\hat{\lambda}_d$ is an average of

$$\frac{M\hat{S}E(\hat{g}_{d,SYN})}{M\hat{S}E(\hat{g}_{d,SYN}) + M\hat{S}E(\hat{g}_d)}$$

over a domain size class.

If there are no poor in a domain, the default estimator is calculated from the whole country, the synthetic estimator uses predictions from the country and composite estimator equals the synthetic one.

Our experiments imply that poverty gap is the most difficult poverty indicator to

estimate, considering the large RRMSE of all estimators. Table 12 shows an experiment with a lot of auxiliary information. All poverty gap estimators, even the default estimator have design bias in small domains, probably due to the non-linear formulation of the indicator. The ordinary predictor (31) is far too biased to be useful. The expanded predictor and corresponding composite estimator are better than the default estimator especially in small domains.

Table 12. Poverty gap estimators assisted by a linear mixed model.
Design: SRSWOR
Quantitative x: educ-thh, empmohh.
Qualitative x: house ownership, lfs-code, socstrat.
Domains: NUTS3 by age by gender (70 domains).
Mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme.

Estimator	BIAS			ARB (%)			RRMSE (%)		
	minor	medium	major	minor	medium	major	minor	medium	major
Default	2.1	0.9	0.4	12.14	4.37	1.78	65.85	43.58	27.26
Predictor	–6.8	–9.8	–14.6	40.09	43.36	57.47	61.49	57.09	62.09
Expanded predictor (17)	–3.1	–3.0	–3.6	17.01	19.61	16.58	23.85	25.43	22.92
Composite	–1.7	–2.1	–2.5	10.91	14.41	11.90	25.63	22.39	18.63

The amount of auxiliary data seems to have an effect on the poverty gap estimation results: in Table 13 involving less auxiliary data than Table 12, the expanded predictor and the frequency-calibrated poverty gap estimator are significantly better than the default estimator only in the smallest domains (expected sample size smaller than 50). Moreover, they are severely biased. The corresponding composite estimators perform better, also in the large domains. Some composite estimators could not be calculated due to limited time. All estimators except the ordinary predictor are robust. Actually, contamination often seemingly improves the properties of estimators**.**

Table 13. Poverty gap in contaminated data.
Design: SRSWOR or PPS by socio-economic status.
Qualitative x-variables: age and gender with interactions, lfs-code and socstrat.
Domains: NUTS3 by gender and age class (70 domains).
Mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme.

Estimator	ARB (%)				RRMSE (%)			
	minor	medium	major	all	minor	medium	major	all
No contamination, SRSWOR								
Default	13.15	5.14	2.07	7.30	66.91	44.17	27.57	48.50
Expanded predictor (18)	45.85	40.04	44.42	43.11	51.91	43.92	47.69	47.64
Composite	28.58	24.24	22.33	25.35	43.28	32.65	29.22	35.66
Predictor	49.85	56.74	62.77	55.66	80.02	75.73	73.33	76.71
n-calibrated predictor	42.42	36.58	39.13	39.25	64.08	48.56	48.45	54.08
n-calibrated composite	23.34	21.74	19.72	21.85	47.37	34.83	29.17	38.02
No contamination, PPS								
Default	13.54	7.66	2.30	8.61	69.74	45.85	28.18	50.60
Expanded predictor (18)	45.03	40.06	45.92	43.09	52.50	45.41	48.96	48.70
Predictor	52.61	52.73	53.76	52.91	67.09	63.09	64.83	64.89
n-calibrated predictor	42.83	37.53	45.84	41.20	59.99	47.18	51.25	52.63
OCAR-CCAR 1 %, PPS								
Default	13.11	7.50	2.06	8.34	69.24	45.68	28.17	50.34
Expanded predictor (18)	44.83	39.87	45.99	42.95	52.62	45.43	48.97	48.76
Predictor	55.16	56.38	57.39	56.16	69.04	66.34	67.88	67.64
n-calibrated predictor	42.51	37.28	45.48	40.91	59.79	47.04	50.97	52.44
OCAR-CCAR 15 %, SRSWOR								
Default	9.68	6.92	4.20	7.28	59.46	41.08	27.71	44.59
Expanded predictor (18)	41.61	35.35	40.77	38.83	52.33	41.82	45.51	46.42
Composite	25.59	20.18	19.37	21.93	41.75	30.27	27.26	33.68
Predictor	92.76	94.22	95.28	93.94	103.19	101.02	99.66	101.49
n-calibrated predictor	41.18	34.02	37.06	37.27	62.97	46.69	46.28	52.41
n-calibrated composite	23.19	18.85	17.59	20.11	45.43	32.37	27.38	35.90
OCAR-NCAR 15 %, PPS								
Default	10.57	6.82	5.48	7.87	64.08	42.27	27.75	46.95
Expanded predictor (18)	34.45	30.56	36.93	33.31	53.01	43.52	46.25	47.50
Predictor	99.27	99.38	99.42	99.35	99.52	99.55	99.58	99.55
n-calibrated predictor	34.57	29.55	37.48	33.04	59.73	45.36	48.22	51.11

Table 14 shows poverty gap estimation results in Amelia data. Here the expanded predictor yields better results than the default method in all domain size classes, since all domains are fairly small (n=2000).

Table 14. Poverty gap in contaminated Amelia data under SRSWOR.
Qualitative x-variables: age and gender with interactions, ISCED, activity and degree of urbanisation.
Domains: age by gender by NUTS2.
Mixed model with NUTS2 random intercepts was fitted to log(income+1) by lme without design weights.

Estimator	ARB (%)				RRMSE (%)			
	minor	medium	major	all	minor	medium	major	all
No contamination								
Default	6.52	3.32	2.38	5.46	51.76	43.77	38.48	48.78
Expanded predictor (18)	18.08	23.77	22.01	19.59	44.38	37.80	30.67	41.62
Composite	10.84	13.83	13.02	11.65	35.84	30.65	25.57	33.73
n-calibrated predictor	14.10	18.40	20.79	15.65	62.97	51.17	41.40	58.37
n-calibrated composite	9.69	9.54	11.04	9.81	43.05	36.48	30.62	40.44
OCAR-CCAR 1 %								
Default	6.24	3.00	2.39	5.20	51.60	43.69	38.44	48.66
Expanded predictor (18)	17.12	23.00	21.29	18.69	44.44	37.58	30.41	41.60
Composite	10.46	13.55	12.83	11.31	35.92	30.59	25.52	33.76
n-calibrated predictor	14.34	17.89	20.08	15.65	63.33	51.27	41.48	58.64
n-calibrated composite	9.89	9.43	11.04	9.93	43.25	36.50	30.73	40.59

The simulation-based method (23) yields fairly good poverty gap estimates, although there seems to be systematic bias: estimates are too large in small domains and too small in large domains (Table 15). As a result, the poverty gap differences between domain size classes apparent in estimation by the default method are not seen in estimates based on the simulation-based method.

Table 15. Poverty gap estimation by the method of Molina and Rao (2010).
Design: SRSWOR.
Quantitative x: educ-thh, empmohh.
Qualitative x: house ownership, lfs-code, socstrat.
Domains: NUTS3 by age by gender
Mixed model with NUTS3 random intercepts was fitted by nlme.

Estimator	BIAS			ARB (%)			RRMSE (%)		
	minor	medium	major	minor	medium	major	minor	medium	major
Simulation-based	2.42	–0.41	–3.59	35.96	19.14	13.51	41.28	24.96	17.77
Default	0.72	1.02	0.37	10.09	4.82	1.85	69.66	44.18	27.54

Although these results are promising, experiments with Gini coefficient and quintile share were disappointing due to large bias. The distribution of the equivalized incomes differs from assumed log-normal distribution: there are fewer rich people

than expected. As a consequence, some of the simulated incomes were unrealistically large. However, in other countries, the distribution of equivalized incomes may be closer to log-normal, and then the method of Molina and Rao is probably the best method available, if minimization of MSE is required. Better results might also be obtained with a more realistic income distribution.

Table 16 compares two bootstrap techniques used in estimating the MSE of the synthetic component in a composite estimator. K=500 samples were used in the bootstrap and RAST correction was applied. Estimating the MSE of the synthetic component in the composite estimator by parametric bootstrap may yield small benefits over the simple equation (21), but it requires much more computing time.

Table 16. Composite estimates (32) of poverty gap with MSE of synthetic component estimated by ordinary bootstrap (21) or by parametric bootstrap (22).
Design: PPS by education level.
Quantitative x: educ-thh, empmohh.
Qualitative x: house ownership, lfs-code, socstrat.
Domains: NUTS3 by age by gender.
A mixed model with NUTS3 random intercepts was fitted by nlme without using design weights.

Bootstrap method	ARB (%)			RRMSE (%)		
	minor	medium	major	minor	medium	major
ordinary bootstrap	11.30	14.76	12.22	25.65	22.64	18.63
parametric bootstrap	11.25	13.98	12.56	25.22	22.60	18.69

4.6 Quintile share ratio S20/S80

S20/S80 ratio, or *quintile share ratio*, is the ratio of the average income of the poorest 20% of people (first quintile) to the average income of the richest 20% of people (fifth quintile). To find the first quintile, we sort the persons by income. The first quintile $q_{d,20}$ is the set of poorest people in domain *d* whose sum of weights is just below or at 20% of the total sum of weights. The *default* (direct) *estimator* of S20 in domain *d* is the Hájek estimator $\hat{S}20_{d;HT}$ of the mean income in the first sample quintile, that is,

the HT estimate of the first quintile total income divided by the estimated population size of the quintile:

$$\hat{S}20_{d;Hajek} = \frac{\sum_{k \in q_{d,20}} a_k y_k}{\sum_{k \in q_{d,20}} a_k}.$$

Similarly, the fifth quintile $q_{d,80}$ is the set of domain's richest people with sum of weights just below or at 20% of the total of weights. The S80 estimate is defined as

$$\hat{S}80_{d;Hajek} = \frac{\sum_{k \in q_{d,80}} a_k y_k}{\sum_{k \in q_{d,80}} a_k}$$

and the direct quintile share estimate is

$$\hat{q}_{d;Hajek} = \frac{\hat{S}20_{d;Hajek}}{\hat{S}80_{d;Hajek}}. \qquad (33)$$

For the synthetic estimators of S20 and S80 in domain *d*, the quintiles $q_{SYN;d,20}$ and $q_{SYN;d,80}$ are defined in population domain as if the weights were constant. The *synthetic estimator* of S20 is the average of predictions $\hat{y}_k$ over the first quintile $q_{SYN;d,20}$:

$$\hat{S}20_{d;SYN} = \frac{\sum_{k \in q_{SYN;d,20}} \hat{y}_k}{\sum_{k \in U_d} I\{k \in q_{SYN;d,20}\}}.$$

The synthetic quintile share estimator is

$$\hat{q}_{d;SYN} = \frac{\hat{S}20_{d;SYN}}{\hat{S}80_{d;SYN}}. \qquad (34)$$

It is also possible to estimate the quintile share using an estimated Lorenz curve:

$$\hat{q}_{d;HT} = \frac{L_{HT;d}(0.2)}{1 - L_{HT;d}(0.8)}$$

and

$$\hat{q}_{d;SYN} = \frac{L_{SYN;d}(0.2)}{1 - L_{SYN;d}(0.8)}.$$

These estimators have yielded similar results as the estimators (32) and (33).

Composite estimator (20) of the quintile share ratio for domain *d* is given by

$$\hat{q}_{d;COMP} = \hat{\lambda}\hat{q}_d + (1 - \hat{\lambda})\hat{q}_{d,SYN}, \qquad (35)$$

where $\hat{\lambda}$ was constructed similarly as in (32).

Default estimates from the smallest domains with at most one observation are replaced by default estimates from the whole country.

Table 17 shows experimental results with quintile share estimators assisted by a linear fixed-effects model. The ordinary predictor (34) is definitely design biased. The expanded predictor yields much better results than the default estimator in all domain size classes. It does not have much design bias.

Table 17. Quintile share estimators assisted by a linear fixed effects model.
Design: SRSWOR.
Quantitative x: educ-thh, empmohh.
Qualitative x: house ownership, lfs-code, socstrat.
Domains: 36 NUTS4 regions.
Model was fitted to log(income+1).

Estimator	BIAS			ARB (%)			RRMSE (%)		
	minor	medium	major	minor	medium	major	minor	medium	major
Default estimator	0.6	0.3	0.2	1.88	1.12	0.59	18.01	13.80	9.19
Predictor	13.2	13.5	12.8	44.63	45.47	45.49	44.95	45.78	45.81
Expanded predictor (17)	0.8	–0.2	1.4	5.63	4.18	6.17	6.25	5.11	6.88
Composite	0.7	0.0	1.0	4.57	3.22	4.27	7.22	5.53	6.14

Tables 18-20 summarize our experiments with contaminated data under SRSWOR. The expanded predictor and frequency-calibrated predictor have the smallest RRMSE and not too much design bias. Moreover, they are more robust than the default estimator. Composite estimators suffer from bias in contaminated data.

Table 18. Quintile share in contaminated data (Finnish data set)
Design: SRSWOR.
Qualitative x-variables: age and gender with interactions, lfs-code and socstrat.
Domains: NUTS4.
Mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme.

Estimator	ARB (%)				RRMSE (%)			
	minor	medium	major	all	minor	medium	major	all
No contamination								
Default	2.31	1.23	0.57	1.14	18.17	13.77	9.17	12.72
Expanded predictor (18)	2.75	4.47	8.65	5.74	4.06	5.76	9.38	6.83
Composite	2.23	3.48	5.51	4.04	6.03	5.87	7.35	6.43
n-calibrated predictor	5.61	5.00	9.22	6.61	8.67	7.19	10.20	8.48
n-calibrated composite	4.85	3.97	5.78	4.74	8.77	6.75	7.68	7.36
OCAR-CCAR 1 %								
Default	11.33	13.96	15.12	14.02	27.92	24.41	19.55	23.14
Expanded predictor (18)	8.52	10.25	4.63	7.98	9.06	10.79	5.99	8.82
Composite	9.09	10.88	7.60	9.45	12.74	13.07	9.84	11.86
n-calibrated predictor	3.67	7.93	4.31	6.03	7.89	9.58	6.23	8.14
n-calibrated composite	5.86	9.03	7.41	8.01	11.69	12.01	9.82	11.17
OCAR-NCAR 1 %								
Default	31.91	49.01	80.10	57.87	59.10	70.22	87.79	75.02
Expanded predictor (18)	10.84	12.02	5.31	9.43	11.70	13.10	7.49	10.88
Composite	20.69	32.47	62.62	41.72	32.01	43.09	68.66	50.79
n-calibrated predictor	5.80	9.56	4.76	7.30	9.54	11.52	7.48	9.79
n-calibrated composite	17.85	30.86	62.40	40.44	30.27	41.72	68.46	49.78
OAR-CAR								
Default	35.59	50.82	67.85	54.85	58.37	65.41	71.64	66.68
Expanded predictor (18)	17.08	18.09	9.31	14.78	17.39	18.44	10.21	15.32
Composite	25.90	36.56	55.20	41.81	35.02	43.57	58.21	47.67
n-calibrated predictor	12.27	15.47	8.65	12.56	14.46	16.69	10.00	13.97
n-calibrated composite	23.45	35.40	55.16	40.88	33.33	42.68	58.20	46.98

Table 19. Unit-level quintile share estimators in contaminated data (Amelia).
Design: SRSWOR.
Qualitative x-variables: age and gender with interactions, ISCED, activity and degree of urbanisation.
Domains: DIS regions.
Mixed model with DIS random intercepts was fitted to log(income+1) by nlme.

Estimator and contamination model	ARB (%)				RRMSE (%)			
	minor	medium	major	all	minor	medium	major	all
No contamination								
Direct	4.9	4.6	3.4	4.4	43.5	41.7	38.5	41.3
Expanded predictor	12.3	8.6	5.7	8.9	16.0	13.6	11.4	13.7
Composite	9.8	7.1	4.7	7.2	16.0	14.6	12.6	14.5
OCAR-CCAR 1%								
Direct	7.9	9.1	10.8	9.2	43.8	41.8	39.3	41.7
Expanded predictor	14.3	8.5	5.7	9.5	18.1	14.2	12.2	14.8
Composite	12.8	8.0	6.9	9.2	18.8	15.9	14.0	16.2
OCAR-NCAR 1%								
Direct	9.1	12.3	16.7	12.6	53.3	53.2	53.2	53.2
Expanded predictor	15.0	8.9	6.6	10.1	18.6	14.5	12.4	15.1
Composite	13.4	9.4	9.3	10.6	21.3	19.3	18.6	19.7

Table 20. Quintile share estimators with aggregated auxiliary data in contaminated data (Amelia).
Design: SRSWOR.
Qualitative x-variables: age and gender with interactions, ISCED, activity and degree of urbanisation.
Domains: DIS regions.
Mixed model with DIS random intercepts was fitted to log(income+1) by nlme.

Estimator and contamination model	ARB (%)				RRMSE (%)			
	minor	medium	major	all	minor	medium	major	all
No contamination								
Direct	4.9	4.6	3.4	4.4	43.5	41.7	38.5	41.3
n-calibrated predictor	11.1	13.3	10.6	11.9	31.3	29.6	25.9	29.1
n-calibrated composite	8.8	10.8	8.9	9.7	27.9	26.6	23.5	26.1
OCAR-CCAR 1%								
Direct	7.9	9.1	10.8	9.2	43.8	41.8	39.3	41.7
n-calibrated predictor	10.9	10.3	7.0	9.6	30.6	27.7	23.7	27.5
n-calibrated composite	9.0	7.0	4.9	7.0	27.2	24.5	21.1	24.4
OCAR-NCAR 1%								
Direct	9.1	12.3	16.7	12.6	53.3	53.2	53.2	53.2
n-calibrated predictor	11.0	9.6	6.3	9.1	30.3	27.1	23.0	26.9
n-calibrated composite	9.4	6.4	4.3	6.7	28.5	26.0	23.3	26.0

Table 21 shows a contamination experiment with PPS. The PPS design seems to result in larger RRMSE of expanded predictor and frequency-calibrated estimator under OCAR-CCAR but other changes are small (compare to Table 18).

Table 21. Quintile share in contaminated data under PPS.
Design: PPS by socio-economic status.
Qualitative x-variables: age and gender with interactions, lfs-code and socstrat.
Domains: NUTS4.
Mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme without using the design weights.

Estimator	ARB (%)				RRMSE (%)			
	minor	medium	major	all	minor	medium	major	all
No contamination								
Default	3.13	1.69	0.66	1.52	20.66	15.58	9.86	14.22
Expanded predictor (18)	2.86	4.57	8.71	5.83	3.96	5.71	9.34	6.78
Composite	2.39	3.47	5.69	4.12	7.20	6.22	7.61	6.86
n-calibrated predictor	6.27	5.06	9.07	6.68	9.27	6.88	9.96	8.32
n-calibrated composite	5.43	3.99	5.86	4.86	9.95	6.86	7.87	7.65
OCAR-CCAR 1 %								
Default	10.52	13.62	14.99	13.69	29.05	25.30	20.02	23.91
Expanded predictor (18)	7.76	9.61	4.62	7.55	8.39	10.23	6.00	8.45
Composite	8.18	10.20	7.06	8.78	13.01	12.68	9.69	11.65
n-calibrated predictor	3.51	7.49	4.56	5.88	7.99	9.00	6.30	7.88
n-calibrated composite	4.65	8.46	6.94	7.38	12.19	11.68	9.71	11.04
OCAR-NCAR 1 %								
Default	29.55	49.00	80.02	57.50	58.55	70.38	87.64	74.97
Expanded predictor (18)	16.80	17.99	9.87	14.89	19.85	21.12	15.06	18.75
Composite	22.38	34.98	64.44	43.87	35.20	45.60	70.37	53.10
n-calibrated predictor	10.81	15.20	9.47	12.52	16.27	19.05	14.94	17.18
n-calibrated composite	19.52	33.57	64.40	42.75	33.24	44.39	70.35	52.22
OAR-CAR								
Default	27.28	42.44	63.48	47.93	53.99	61.18	69.71	63.26
Expanded predictor (18)	16.15	17.04	8.50	13.83	16.52	17.48	9.51	14.47
Composite	21.04	30.21	48.92	35.69	31.39	38.28	53.48	42.81
n-calibrated predictor	10.43	14.45	8.05	11.58	13.59	15.69	9.46	13.15
n-calibrated composite	18.05	28.79	48.83	34.54	29.44	37.07	53.43	41.92

4.7 Classifying domains by poverty

The estimated indicators are probably used in decision making. Thresholds for a poverty indicator have been used in regional allocation of resources (e.g. Zaslavsky and Schirm, 2002, and others in Journal of Official Statistics, vol. 18, no. 3). As an application of a poverty indicator, domains might be classified as poor (“positive”) and not poor (“negative”) using a threshold. Large values of poverty rate, poverty gap and Gini coefficient, or small values of quintile share imply poverty. Ranking domains by poverty indicator may identify domains with greatest problems. For example, we classify a domain as poor, if its rank by quintile share is small.

In the classification terminology, a domain is called true positive if it is correctly classified as positive (poor), and true negative if it is correctly classified as negative. A truly positive domain is positive in truth. Precision (positive predictive value) is the ratio of the number of true positives to the number of all positive classifications. It estimates the probability that a domain classified as poor is poor in truth. Sensitivity (recall, true positive rate) is the ratio of the number of true positives to the number of truly positive domains. This can be interpreted as the probability of classifying correctly a truly poor domain. Accuracy is the proportion of correct classifications, composed of true positives and true negatives. These measures are calculated in separate size-classes, as averages over all simulations. For example, precision in the small size class in a single simulation is the proportion of true positive small domains of all positively classified small domains in the simulation.

Table 22 compares poverty rate estimators' ability to classify domains to classes by poverty rate over 0.2 (positive domains) or under 0.2 (negative domains). EBP(Y) seems to have the best overall accuracy but it does not identify well domains that are deemed positive by the fixed threshold.

Table 22. Success of poverty rate estimators in identifying 7 poorest domains.
Design: PPS by socio-economic status.
Qualitative x-variables: age and gender with interactions, lfs-code and socstrat.
Domains: NUTS3 by gender and age class.
Mixed model with NUTS3 random intercepts was fitted to log(income+1) by nlme without using design weights.

	PRECISION (%)			SENSITIVITY (%)			ACCURACY (%)		
Estimator	**minor**	**medium**	**major**	**minor**	**medium**	**major**	**minor**	**medium**	**major**
a) Classification by fixed threshold (0.2)									
Default	50.7	32.4	.	59.8	49.4	.	83.5	83.9	93.1
MLGREG	50.0	32.6	.	59.7	50.4	.	83.3	83.7	93.0
EBP(Y)	37.8	35.8	.	21.2	29.0	.	86.4	90.3	97.1
b) Classification by rank									
Default	55.5	37.2	.	51.3	35.8	.	85.4	87.6	97.4
MLGREG	55.8	37.2	.	51.0	36.1	.	85.3	87.7	97.4
EBP(Y)	78.3	51.7	.	40.6	52.1	.	88.6	88.4	91.4

In our experiments, the expanded predictors of quintile share, poverty gap and Gini coefficient had the best accuracy in classification of small domains by rank. However, the default estimator had the best overall accuracy in classification by rank. No clear picture emerged from classification by threshold under SRSWOR and OCAR-CCAR (1%): the best classifiers for each poverty indicator were EBP(Y) for poverty rate, frequency-calibrated estimator for quintile share, the default estimator for Gini coefficient and the expanded predictor for poverty gap.

We expected that small design bias is important in identifying poor domains given a fixed threshold and small RRMSE is important in classification by rank. Table 22 gives some support to these expectations. In our experiments, the good accuracy of classification by rank with EBP(Y), expanded predictor and the frequency-calibrated estimator in small domains is probably due to their small RRMSE. A more complete picture of the classification abilities of estimators would be obtained by studying accuracy over a range of thresholds and ranks.

5 Case study: Estimation of poverty rate and its variance

5.1 Introduction

This section:

1) Compares model assisted generalized regression estimators (GREG), dampened regression estimator (DRE), and model dependent pseudo SYN and EBLUP estimators for poverty estimation in domains under stratified sampling, and
2) Studies the goodness of the Sen-Yates-Grundy (SYG), bootstrap, and augmented SYG variance estimators for the above mentioned poverty rate estimators.

5.2 Design

Population

SRSWOR sample of 20,000 subjects from population of N = 1,000,000 used in D2.1.A.

Domains

D = 30 domains constructed from age (3 categories), sex (2 categories), and NUTS3 (5 categories). Domains are mutually exclusive and exhaustive, unplanned (that is domain sample sizes n_d are random), and may cut across the strata.

Target variable

Poverty rate in domains, defined as #poor/domain size. Poverty indicator is 1 if equivalized household income is less than 60% of the median of equivalized household income. The median is estimated for each sample, and the poverty indicator is based on the estimated median.

Sampling design

Stratified sampling with simple random sampling without replacement within strata; number of samples = 1,000. Table 23 shows the sample sizes by strata.

Table 23. Sampling scheme (Stratified, SRSWOR within strata)

Education (stratum)	*N*	*n*	**Sampling fraction** (%)
0 (lowest)	11381	1419	12.5
3	4246	353	8.3
4	2633	164	6.2
5	769	32	4.2
6	340	11	3.2
7	568	19	3.3
8 (highest)	63	2	3.2
Total	20000	2000	10.0

The expected sample size per domain is 66.7, with minimum 18.5 and maximum 121.7. Domains are categorized into minor, medium and major according to the expected sample size as shown in Table 24.

Table 24. Expected sample size $E(n_d)$ by domain type

	Minor domain $E(n_d)$ 18.5-49	Medium domain $E(n_d)$ 50-99	Major domain $E(n_d)$ 100+
Number of domains	10	12	8
Average E(nd)	27.3	69.2	112.1
Min E(nd)	18.5	53.4	104.2
Max E(nd)	38.9	92.8	121.7

Mathematical notation

Table 25 shows the notation used in this paper. With this notation, poverty rate in domain d (P_d) can be expressed in three convenient forms:

$$P_d = \frac{1}{N_d}\sum_{U_d} y_i = \frac{1}{N_d}\sum_{U} y_{id} = \frac{1}{N_d}\sum_{h=1}^{7}\sum_{U_h} y_{id} \ . \qquad (36)$$

Table 25. Summary of notation

Symbol	Description
U, U_d	Population; population in domain d
N, N_d	Population size; population size in domain d
s, s_d	Sample set, sample set in domain d
i	Index for an individual ($i = 1, 2, ..., 20000$)
h	Stratum index ($h = 1, 2, ..., 7$)
d	Domain index ($d = 1, 2, ..., 30$)
n, n_d	Sample size; sample size in domain d
y_i	Poverty indicator for individual i: $y = 1$ if poor, 0 otherwise
y_{id}	Domain poverty indicator: $y_{id} = y_i$ in domain d, zero otherwise
T_d	Total of y in domain d
P_d	Poverty rate in domain d (the variable to be estimated)
w_i	Sampling weight

We build the estimators on the right hand side expression of the poverty rate. The familiar Horwitz-Thompson ($\hat{P}_{d,HT}$) estimator for P_d is

$$\hat{P}_{d,HT} = \frac{1}{N_d}\sum_{h=1}^{7}\hat{T}_{h,d}, \quad \text{where} \quad \hat{T}_{h,d} = \sum_{i \in s_h} w_i y_{id}. \tag{37}$$

In the Horvitz-Thompson estimator above, we first estimate the stratum totals for domain d, then sum these stratum totals, and finally divide by domain size to obtain the poverty estimate.

5.3 Estimators

5.3.1 Poverty rate estimators

We study the properties of four poverty rate estimators: the generalized regression estimator (GREG; Särndal et al. 1992; Lehtonen and Veijanen 1998), the dampened regression estimator (DRE; Särndal and Hidiroglou 1989); pseudo-synthetic estimator (SYN; You and Rao 2002) and pseudo-EBLUP estimator (EBLUP; You and Rao 2002). The SYN and EBLUP estimators are called pseudo-SYN and pseudo-EBLUP as we use weights when estimating the models.

All estimators are built using the principle shown in equation 36: first the stratum totals are estimated, then they are summed over strata, and finally divided by domain size. Thus all four estimators look like this:

$$\hat{P}_d = \frac{1}{N_d}\sum_{h=1}^{7}\hat{T}_{h,d} \tag{38}$$

Only $\hat{T}_{h,d}$ (and the model) differentiate the estimators. Table 26 shows $\hat{T}_{h,d}$ for the four estimators considered, and Table 27 the models that are used in conjunction with the four estimators.

Table 26. Summary of estimators

Estimator	$\hat{T}_{h,d}$
GREG	$\hat{T}_{h,d}^{GREG} = \sum_{i\in U_h}\hat{y}_{id} + \sum_{i\in s_h} w_i\left(y_{id} - \hat{y}_{id}\right)$
DRE	$\hat{T}_{h,d}^{DRE} = \sum_{i\in U_h}\hat{y}_{id} + \hat{\lambda}_d\sum_{i\in s_h} w_i\left(y_{id} - \hat{y}_{id}\right),\quad \hat{\lambda}_d = \left(\hat{N}_d / N_d\right)^{c-1},\quad c = \begin{cases} 0 \text{ if } \hat{N}_d \geq N_d \\ 2 \text{ if } \hat{N}_d < N_d. \end{cases}$
SYN	$\hat{T}_{h,d}^{SYN} = \sum_{i\in U_h}\hat{y}_{id}$
EBLUP	$\hat{T}_{h,d}^{EBLUP} = \sum_{i\in U_h - s_h}\hat{y}_{id} + \sum_{i\in s_h} y_{id}$

Table 27. Models used in the estimators shown in Table 26

Estimator	Model number and description	Domain intercepts	Intercepts random/fixed	Variables used in all models
GREG	1. Linear, no domain int.	No	-	Sex Own house LFS status (3 cat) Age (5 cat)
	2. Linear with domain int	Yes	Fixed	
	3. Logistic, no domain int.	No	-	
	4. Logistic with domain int.	Yes	Fixed	
	5. Linear random int. model	Yes	Random	
DRE	1. Linear, no domain int.	No	-	
	2. Linear with domain int	Yes	Fixed	
	3. Logistic, no domain int.	No	-	
	4. Logistic with domain int.	Yes	Fixed	
	5. Linear random int. model	Yes	Random	
SYN	5. Linear random int. model	Yes	Random	
EBLUP	5. Linear random int. model	Yes	Random	

5.3.2 Variance estimators

As with the poverty rate, we estimate the variances by first estimating the stratum specific variance components, then summing these up, and finally scaling appropriately. The variance estimators we use are the SYG variance estimator; without replacement bootstrap; and augmented SYG variance estimator. These are implemented as follows.

1. The standard Sen-Yates-Grundy (SYG) type variance estimator is based on the model residuals (Särndal et al. 1992). The variance is estimated for each strata and summed up to the population level. More specifically, the variance for $\hat{P}_d$ is estimated as

$$V_{SYG}\left(\hat{P}_d\right)=\left(\frac{1}{N_d}\right)^2\sum_{h=1}^{7}V\left(\hat{T}_{h,d}\right), \qquad (39)$$

where

$$V_{SYG}\left(\hat{T}_{h,d}\right)=\frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\sum_{s_h}\left(e_{id}-\overline{e}_{dh}\right)^2, \quad e_{id}=y_{id}-\hat{y}_{id}, \text{ and } \overline{e}_{dh}=\frac{1}{n_h}\sum_{s_h}e_{id}. \qquad (40)$$

Note that e_{id} is zero outside the domain, and that sampling weights are not needed in (40) because the weights are constant within strata.

2. Bootstrap without replacement (Efron 1979, Särndal et al. 1992; Booth et al. 1994). The bootstrap procedure is implemented as follows.

a. Generate a bootstrap population (N = 20,000) by drawing a stratified with replacement sample from the original sample using the inverses of the original sampling fractions. The bootstrap population has the same stratum sizes as the original population, and each unit in each stratum in the bootstrap population belongs to the same stratum in the original population.

b. Use the original sampling scheme (without replacement stratified sampling) to draw a bootstrap sample from the bootstrap population

c. Calculate the poverty estimates for each domain and for each estimator

d. Repeat b.-c. 200 times, and calculate the variance estimate as the variance of the 200 pseudo-estimates

3. Augmented SYG estimator (Myrskylä 2007). To appreciate this estimator, note first that in the SYG estimator the terms $e = y - \hat{y}$ (subscripts dropped for clarity) are

sample fit residuals which aim to estimate $E = y - y^p$, the population fit residuals (y^p denotes the prediction obtained using the whole population to estimate the model). The population fit residuals can be decomposed as

$$E = y - y^p = \left(y - \hat{y}\right) + \left(\hat{y} - y^p\right) = \hat{e} + \hat{e}^p . \qquad (41)$$

Thus the sample fit residual in the SYG estimator estimates part of the population fit residual, but ignores the uncertainty that comes from the difference between the sample fit residuals and population fit residuals. This can be interpreted also so that the SYG estimator ignores the uncertainty which is due to the randomness in the model parameters.

The augmented SYG estimator (AUG) takes the terms $\hat{e}^p$ into account using a bootstrap-like procedure. The AUG estimator for the stratum *h* total is

$$V_{AUG}\left(\hat{T}_{h,d}\right) = V_{SYG}\left(\hat{T}_{h,d}\right) + V_A\left(\hat{T}_{h,d}\right), \qquad (42)$$

where $V_A\left(\hat{P}_d\right)$ corrects for the error $\hat{e}^p$ and is estimated as follows:

a. Generate a bootstrap population (N = 20,000) by drawing a stratified with replacement sample from the original sample using the inverses of the original sampling fractions. The bootstrap population has the same stratum sizes as the original population, and each unit in each stratum in the bootstrap population belongs to the same stratum in the original population.

b. Calculate the population fit predictions for the bootstrap population

c. Use the original sampling scheme (without replacement stratified sampling) to draw a bootstrap sample from the bootstrap population

d. Estimate the sample fit model; calculate $\hat{e}^p = \hat{y} - y^p$; and estimate the variance contribution due to $\hat{e}^p$ as

$$V_A^*\left(\hat{T}_{h,d}\right) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\sum_{s_h}\left(\hat{e}_{id}^p - \bar{\hat{e}}_{dh}^p\right)^2, \text{ where } \bar{\hat{e}}_{dh}^p = \frac{1}{n_h}\sum_{s_h}\hat{e}_{id}^p. \quad (43)$$

e. Repeat b.-d. 10 times, and calculate the variance contribution $V_A\left(\hat{T}_{h,d}\right)$ as the average of the estimates (43).

After obtaining the stratum-specific variance estimates for the stratum totals (42), the final variance estimate is obtained as in (39) by summing the variance components and dividing by the square of the domain size.

5.4 Results

5.4.1 Poverty rate estimators

ARB (Absolute Relative Bias). GREG estimators are approximately unbiased (Table 28). Composite estimators (DRE estimators) have slightly larger bias than GREG estimators, but the differences are small. For GREG and DRE estimators, the bias decreases with sample size. SYN and EBLUP estimators have large biases, and the bias does not decrease with domain sample size.

RRMSE (Relative Root Mean Square Error). For GREG estimators, accuracy is almost the same for all models (Table 28). Domain intercepts do not improve accuracy, and GREG-log is not more accurate than GREG-lin.

Table 28. Absolute relative bias and relative root mean square error for the poverty rate estimators in a Monte Carlo simulation with 1,000 replicates

Absolute Relative Bias ARB (%)

	Minor domains $E(n_d)$ 18.5-49				Medium domains $E(n_d)$ 50-99				Major domains $E(n_d)$ 100+			
Model*	GREG	DRE	SYN	EBLUP	GREG	DRE	SYN	EBLUP	GREG	DRE	SYN	EBLUP
1. Lin, no domain int.	2.7	3.5			1.3	1.5			0.9	1.7		
2. Lin with domain int	2.8	2.8			1.3	1.3			1.0	1.0		
3. Log, no domain int.	2.7	3.6			1.3	1.6			1.0	1.7		
4. Log with domain int.	2.9	2.9			1.4	1.4			0.9	0.9		
5. Lin random int. model	2.7	3.5	12.5	11.2	1.3	1.5	6.2	5.8	0.9	1.5	9.9	8.8

RRMSE (%)

	Minor domains $E(n_d)$ 18.5-49				Medium domains $E(n_d)$ 50-99				Major domains $E(n_d)$ 100+			
Model*	GREG	DRE	SYN	EBLUP	GREG	DRE	SYN	EBLUP	GREG	DRE	SYN	EBLUP
1. Lin, no domain int.	41.5	37.1			29.5	27.2			25.9	24.4		
2. Lin with domain int	42.2	42.2			29.7	29.7			26.1	26.1		
3. Log, no domain int.	41.4	37.0			29.0	26.7			25.4	23.9		
4. Log with domain int.	42.3	42.3			29.3	29.3			25.7	25.7		
5. Lin random int. model	41.4	37.3	21.1	21.3	29.5	27.5	14.0	14.5	25.9	24.8	17.6	17.3

*The fixed effects part is "global intercept + house + lfs2 + lfs3 + age1 + age2 + age3 + age4 + sex" + domain intercepts for models 2, 4. Model 5 has random domain intercepts. All models use sampling weights in estimation.

For DRE composite estimators, accuracy improves if domain intercepts are **not** used, or if they are included as random effects (Table 28). If domain intercepts are included as fixed in the model, the estimator is equivalent to the GREG estimator. The fact that accuracy is gained if domain intercepts are not used can be explained as follows: the error correction term $\sum_{s^d} w_i e_i^d$ is approximately zero if domain intercepts are used. Then, the DRE estimator

$$\sum_{U_d} \hat{y}_i^d + \hat{\lambda}^d \sum_{s_d} w_i e_i^d \approx \sum_{U_d} \hat{y}_i^d ,$$

so the DRE composite estimator is equivalent to the GREG estimator, and use of λ-weighted error correction term has no effect on the estimator. If, however, domain

intercepts are **not** used, the error correction term is not zero, and the estimation error in $\hat{\lambda}^d \sum_{s_d} w_i e_i^d$ is negatively correlated with the estimation error in $\sum_{U_d} \hat{y}_i^d$. The negative correlation between these variables results in improvement in accuracy.

For SYN and EBLUP estimators the RRMSE is generally smaller than for DRE and GREG estimators (Table 28). This holds even in domains with expected sample size > 100. SYN and EBLUP, however, are biased (see the ARB table).

5.4.2 Variance estimators

Table 29. Empirical coverage rates with nominal coverage level 95.0 for three variance estimators in a Monte Carlo simulation with 1,000 replicates

		Coverage Rate CR (%) by variance estimator								
		$E(n_d)$ 18.5-49			$E(n_d)$ 50-99			$E(n_d)$ >= 100		
Estimator	**Model**	**SYG**	**BWO**	**AUG**	**SYG**	**BWO**	**AUG**	**SYG**	**BWO**	**AUG**
GREG	1. Lin, no domain int.	92.5	92.7	93.9	93.4	93.6	94.8	93.7	93.8	95.3
	2. Lin with domain int	89.7	90.3	92.8	92.1	92.3	94.4	92.7	92.9	94.8
	3. Log, no domain int.	92.5	92.7	94.0	93.1	93.3	94.9	93.4	93.7	95.2
	4. Log with domain int.	89.6	90.2	93.2	91.2	91.3	93.7	91.9	92.1	94.2
	5. Lin random int. model	92.4	93.0	95.6	93.1	93.5	95.4	93.5	93.9	95.5
DRE	1. Lin, no domain int.	94.7	92.0	95.8	95.0	92.8	96.1	94.8	92.9	95.8
	2. Lin with domain int	89.7	90.3	92.8	92.1	92.3	94.4	92.7	92.9	94.8
	3. Log, no domain int.	94.8	92.1	96.0	94.7	92.5	95.8	94.6	92.6	95.9
	4. Log with domain int.	89.6	90.2	93.2	91.2	91.3	93.7	91.9	92.1	94.2
	5. Lin random int. model	94.4	92.1	96.9	94.5	92.7	96.5	94.4	93.0	96.2
SYN	5. Lin random int. model	95.7	95.7	95.7	97.6	97.6	97.6	85.7	85.7	85.7
EBLUP	5. Lin random int. model	95.5	95.5	95.5	97.0	97.0	97.0	86.8	86.8	86.8

GREG estimators. In minor, medium, and even large domains, both SYG and BWO (bootstrap without replacement) underestimate the variance (Table 29). The errors are larger for smaller domains and if the model has domain intercepts. The differences between SYG and BWO are small, although BWO performs slightly better in most

cases. AUG (Augmented SYG) is most accurate, giving coverage rates close to 95% in domains with expected sample size >= 50. In smaller domains, coverage rates are slightly too small for AUG, but markedly closer to 95.0 than they are for SYG or BWO.

DRE (composite) estimators. For DRE estimators with models 2 and 4, the results are the same as they are with GREG estimators (Table 29). The similarity of the results follows from the fact that is the estimator has a fixed domain intercept, DRE estimator is almost equivalent to GREG estimator. So, for models 2 and 4, SYG and BWO underestimate the variance in all domains, AUG underestimates the variance only in smallest domains and even then the error is smaller than it is for SYG or BWO.

For “true” DRE estimators which have models 1, 3 and 5, the SYG estimator performs very well, even better than BWO or AUG. This is due to errors that cancel each other: SYG underestimates the variance of GREG, but DRE estimator has slightly smaller variance than GREG. In this simulation, these errors happen to cancel out, resulting in coverage rates close to 95%. BWO consistently underestimates the variance, but the errors get smaller in larger domains. AUG performs quite well, but occasionally overestimates the variance. This is because the estimator is built for GREG, and estimates the variance for GREG quite well, but the DRE estimator has slightly smaller variance than GREG.

(Pseudo) SYN and EBLUP estimators. All variance estimators deliver the same coverage rates for these estimators (Table 29). This is because whether or not the confidence interval captures the true value depends on the bias in the estimators, not on the relatively small differences in the variance estimators. On the surface, it looks as if the estimators did a good job in estimating the variance for small domains. This, however, is illusory, since the coverage rates are averaged over several domains, and domain-specific coverage rates range from 80.5 to 99.3% (for GREG and DRE estimators, the average coverage rates reflect accurately the average difference from 95.0%). In medium domains, the coverage rates for SYN and EBLUP are on average too high, and in major domains, the coverage rates are too low. None of the three variance estimators should be recommended for SYN or EBLUP estimators.

6 Discussion of results

6.1 General

Domain size is the most important factor affecting accuracy of estimation in a domain. Absolute bias and RRMSE were largest in small domains. With direct estimators and small samples, the estimates vary greatly, and show too large disparities between domains. On the other hand, differences between synthetic estimates are too small.

Sampling design does not seem to affect estimators much. EBP(Y) tended to have somewhat larger bias with varying probability sampling designs especially when the PPS size variable was not in the used model.

In general, results are not improved by adding domain-specific terms to the used model. We obtained better estimates by including terms such as random intercepts associated with NUTS3 levels when domains were defined by NUTS4, for example.

6.2 New predictors

Use of predictors in estimation of poverty indicators is problematic, as the predictions are required for individuals, whereas the response is a household-level equivalized income and the auxiliary variables include both unit- and household-level variables. Models will not fit the data well, especially with apparently unsatisfactory auxiliary data, such as demographic information. If the poverty was measured differently, it might be easier to predict personal income or calculate household level poverty measures using only household-level auxiliary variables.

Ordinary predictors involve predictions plugged into the default formula in place of genuine observations. These predictors are substantially biased: poverty gaps and Gini

coefficients were too small and quintile shares were too large. Due to the bias, the RRMSE of ordinary predictors were even greater than the RRMSE of corresponding default estimator.

The expanded predictors benefit greatly from the transformation of predictions (Eqs. 17 and 18) bringing the distribution of predictions closer to the distribution of observations. Both bias and RRMSE decreased due to the transformation, as compared with the ordinary predictors. Inclusion of design weights in the technique probably reduced design bias in experiments with PPS. Moreover, the expanded quintile share and Gini predictors were more robust than the default method or the ordinary predictor. As the expansion incorporates percentiles of observations up to 99th percentile, rare outliers occurring with frequency of 1 percent do not affect the expanded predictor too much. When the proportion of outliers was 15 percent, the expanded predictor failed but not as badly as the other estimators. The breakdown point of the estimator can probably be adjusted by changing the range of percentage points used in the transformation (17) or (18). In small domains, the expanded predictor usually had smaller RRMSE although larger bias than the default estimator. In the largest domains, the default estimator may be preferred to the expanded predictor if there are no outliers, but in contaminated data the expanded predictors appear to be better than the default estimator, although the poverty gap is an exception.

In poverty gap estimation, only the left tail of the distribution of predictions contributes to estimates. The expansion method does not seem to work as well as in quintile share and Gini coefficient, where most of the predictions are included in the estimators.

The frequency-calibrated estimator (Eqs. 18 and 19) was not usually as accurate as the expanded predictor with same auxiliary variables. This was expected, as the frequency-calibrated predictor has access only to the domain frequencies of classes of auxiliary variables in the population, not to unit-level information. The estimator appears to have similar robustness properties as the expanded predictor. However, in the case of the poverty gap, the frequency-calibrated method may perform poorly.

A composite estimator consists of a default estimator and corresponding expanded predictor. In the case of no contamination, these estimators had smaller bias than the expanded predictors, but RRMSE was usually slightly larger. If contamination yields bias in the default estimator, composite estimators consequently suffer from bias. Composite estimators of quintile share or Gini coefficient may not be a good choice if some contamination is suspected. However, we might prefer composite poverty gap estimators over predictors.

Variance and MSE estimation has been considered in selected cases only. Pseudoreplication methods such as bootstrap and jackknife provide applicable options for variance and MSE estimation of the alternative estimators of the poverty indicators discussed in this report. For example, bootstrap estimator of the MSE of an expanded predictor or a frequency-calibrated predictor should incorporate fitting a model to each bootstrap sample. A more extensive discussion on variance and MSE estimation is in Bruch, Münnich and Zins (2011).

Modelling quantiles of equivalized income by quantile regression might be a useful component in an estimator of a poverty indicator. Some new theory is required, however.

6.3 Comparison of outlier and contamination mechanisms

Contamination experiments with a small proportion of outliers (1 % or OAR-CAR) are realistic for income data. In these experiments, the poverty rate estimators are fairly robust because outliers with large income do not affect much the median-based poverty threshold estimator. Outliers with large income yield too large Gini coefficients and too small quintile shares. The default estimator and the ordinary predictor of these indicators were sensitive to outliers. The expansion of predictions (Eqs. 17 and 18) reduced the effect of outliers. Contamination model NCAR yielded much larger bias than CCAR. The OAR outlier model had larger impact than OCAR perhaps because of the larger proportion of outliers and location parameter in the contamination of employed people. For some reason, the expanded predictor and

frequency-calibrated predictor of quintile share and Gini coefficient were most sensitive to OAR-CAR whereas the default estimator was most sensitive to OCAR-NCAR. Perhaps the expansion technique (18) incorporating percentiles up to the 99th one provided robustness in the case of OCAR-NCAR with 1 % of outliers but suffered from contamination under OAR-CAR with a larger proportion (2 - 4 %) of outliers among people in workforce.

In the most heavily contaminated data sets, the proportion of outliers was 15 %. All except the poverty gap estimators were then clearly affected. Even the poverty rate estimates were smaller, since the proportion of poor people decreased due to contamination. Poverty rate is somewhat sensitive to a large proportion of outliers. When the contamination is independent of income, the median income of poor people remaining in the contaminated data set does not necessarily deviate much from the median income of the poor in the original data set. Therefore poverty gap estimators are not much affected by CCAR, but theoretically NCAR might cause more changes, although our experiments provided no such evidence.

References

Alfons, A., Templ, M., Filzmoser, P., Kraft, S., Hulliger, B., Kolb, J.-P. and **Münnich, R. (2011a)**: *Report on outcome of the simulation study.*
Research Project Report WP6 – D6.1, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Alfons, A., Templ, M., Filzmoser, P., Kraft, S., Hulliger, B., Kolb, J.-P. and **Münnich, R. (2011b)**: *Report on outcome of the simulation study.*
Research Project Report WP6 – D6.2, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Bates, D. (2011) : *Computational methods for mixed models*. Supplement documentation to lme4 package.
URL http://www.cran.r-project.org

Bjornstad, J. F. (2007) : *Non-Bayesian multiple imputation.* Journal of Official Statistics 23, 433-452.

Bruch, C., Münnich, R. and **Zins, S. (2011)**: *Variance Estimation for Complex Surveys.*
Research Project Report WP3 – D3.1, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Chambers, R. L. and **Dorfman, A. H. (2003)**: *Transformed variables in survey sampling.* Working paper M03/21, Southampton Statistical Sciences Research Institute.

Chandra, H., Salvati, N. and **Chambers, R. (2007)**: *Small area estimation for spatially correlated populations – a comparison of direct and indirect model – based methods.* Statistics in Transition **8**, 331-350.

D'Alo, M., Di Consiglio, L, Falorsi, S. and **Solari, F. (2006)** : *Small area estimation of the Italian poverty rate.* Statistics in Transition 7, 771-784.

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

Estevao, V. M. and **Särndal, C.-E. (2004)**: *Borrowing strength is not the best technique within a wide class of design-consistent domain estimators.* Journal of Official Statistics **20**, 645-669.

EURAREA Consortium (2004): *Project Reference Volume.*
URL www.statistics.uk.gov/eurarea

Fabrizi, E., Ferrante, M. R. and **Pacei, S. (2005)**: *Estimation of poverty indicators at sub-national level using multivariate small area models.* Statistics in Transition **7**, 587-608.

Fabrizi, E., Ferrante, M. R. and **Pacei, S. (2007a)**: *Small area estimation of average household income based on unit level models for panel data.* Survey Methodology **33**, 187-198.

Fabrizi, E., Ferrante, M. R. and **Pacei, S. (2007b)**: *Comparing alternative distributional assumptions in mixed models used for small area estimation of income parameters.* Statistics in Transition **8**, 423-439.

Falorsi, P. D., Orsini, D. and **Righi, P. (2006)**: *Balanced and coordinated sampling designs for small domain estimation.* Statistics in Transition **7**, 805-829.

Federal Committee on Statistical Methodology (1993): *Indirect Estimators in Federal Programs.* U.S. Office of Management and Budget, Statistical Policy Working Paper 21.

Hansen, M. H., Hurvitz, W. N. and **Madow, W. G. (1978)**: *On inference and estimation from sample surveys (with discussion).* Proceedings of the Survey Research Methods Section, American Statistical Association, 82-107.

Hansen, M. H., Madow, W. G. and **Tepping, B. J. (1983)**: *An evaluation of model-dependent and probability-sampling inferences in sample surveys (with discussion).* Journal of the American Statistical Association **78**, 776-807.

Haslett, S.J., Isidro, M.C. and **Jones, G. (2010)**: *Comparison of survey regression techniques in the context of small area estimation of poverty.* Survey Methodology **36**, 157-170.

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

Hulliger, B. and **Schoch, T. (2010)**: *Outlier contamination models and simulation schemes.* AMELI Working Paper, June 3, 2010.

Jiang, J. and **Lahiri, P. (2006)**: *Mixed model prediction and small area estimation.* Sociedad de Estadistica e Investigacion Operative Test **15**, 1-96.

Judkins, D. R. and **Liu, J. (2000)**: *Correcting the bias in the range of a statistic across small areas.* Journal of Official Statistics **16**, 1-13.

Kott, P.S. (2009): *Calibration weighting: combining probability samples and linear prediction models.* In: C. R. Rao and D. Pfeffermann (eds.): Handbook of statistics, vol. 29(B). Sample surveys: theory, methods and inference. Elsevier.

Laaksonen, S. (2002): *Traditional and new techniques for imputation.* Statistics in Transition **5**, 1013-1035.

Lehtonen, R. and **Pahkinen, E. (2004)**: *Practical Methods for Design and Analysis of Complex Surveys* (2nd ed.). John Wiley & Sons, Chichester, UK.

Lehtonen, R., Myrskylä, M., Särndal, C.-E. and **Veijanen, A. (2007)**: *Estimation for domains and small areas under unequal probability sampling.* Invited paper, the SAE2007 Conference, Pisa, September 2007. (CD rom).

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

Lehtonen, R., Särndal, C.-E. and **Veijanen, A. (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., Särndal, C.-E. and **Veijanen, A. (2009)**: *Model calibration and generalized regression estimation for domains and small areas.* Invited paper, the SAE2009 Conference, Elche, June/July 2009.

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

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

Lehtonen, R. and **Veijanen, A. (2009)**: *Design-based methods of estimation for domains and small areas.* In: C. R. Rao and D. Pfeffermann (eds.), Handbook of statistics, vol. 29(B). Sample surveys: theory, methods and inference. Elsevier.

Leiten, E. and **Traat, I. (2006)**: *Variance of Laeken indicators in complex surveys.* Tallinn: Statistical Office of Estonia.

Molina, I. and **Rao, J. N. K. (2010)**: *Small area estimation of poverty indicators.* The Canadian Journal of Statistics **38**, 369-385.

Münnich, R. and **Wiegert, R. (2001)**: *The DACSEIS Project.* DACSEIS research paper series No. 1. Research Project IST-2000-26057 DACSEIS.
URL http://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/DRPS1.pdf

Myrskylä, M. (2007): *Generalized regression estimation for domain class frequencies.* Helsinki: Statistics Finland, Research Reports **247**. (PhD dissertation in Statistics, University of Helsinki).

Paddock, S. M., Ridgeway, G., Lin, R. and **Louis, T. A. (2006)**: *Flexible distributions for triple-goal estimates in two-stage hierarchical models.* Computational Statistics & Data Analysis **50**, 3243-3262.

Purcell, N. J. and **Kish, L. (1980)**: *Postcensal estimates for local areas (or domains).* International Statistical Review **48**, 3-18.

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

Rubin, D. (1987): *Multiple imputation for nonresponse in surveys.* John Wiley & Sons, New Jersey.

Schafer, J.L. (1997): *Analysis of Incomplete Multivariate Data*. Chapman & Hall, New York.

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

Shen, W. and **Louis, T. A. (1998)**: *Triple-goal estimates in two-stage hierarchical models.* Journal of the Royal Statistical Society **B 60**, 455-471.

Singh, M. P., Gambino, J. and **Mantel, H. J. (1994)**: *Issues and strategies for small area data.* Survey Methodology **20**, 3-14.

Singh, A. C. and **Mohl, C. A. (1996)**: *Understanding calibration estimators in survey sampling.* Survey Methodology **22**,107-115.

Srivastava, A. K. (2009): *Some aspects of estimating poverty at small area level.* J. Indian Soc. Agric. Stat. **63(1)**, 1-23.

Särndal, C.-E. (1984): *Design-consistent versus model-dependent estimation for small domains.* Journal of the American Statistical Association **79**, 624-631.

Särndal, C.-E. (2007): *The calibration approach in survey theory and practice.* Survey Methodology **33**, 99-119.

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

Torabi, M. and **Rao, J. N. K. (2008)**: *Small area estimation under a two-level model.* Survey Methodology **34**, 11-17.

Veijanen, A. and Lehtonen, R. (2011): *Small Area Estimation of Indicators on Poverty and Social Exclusion. Manual of R codes.*

Research Project Report WP2 – D2.2 Supplement, FP7-SSH-2007-217322 AMELI.
URL http://ameli.surveystatistics.net

Verma, V., Betti, G. and **Gagliardi, F. (2010)**: *Robustness of some EU-SILC based indicators at regional level.* Eurostat Methodologies and Working Papers. Luxembourg: Publications Office of the European Union.

Wu, C. (2003): *Optimal calibration estimators in survey sampling.* Biometrika **90**, 937-951.

Wu, C. and **Sitter, R. (2001)**: *A model-calibration approach to using complete auxiliary information from survey data.* Journal of the American Statistical Association 96, 185-193.

You, Y. and **Rao, J. N. K. (2002)**: *A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights.* The Canadian Journal of Statistics ***30***, *431-439.*

Zaslavsky, A. M. and **Schirm, A. L. (2002)**: *Interactions between survey estimates and federal funding formulas.* Journal of Official Statistics **18**, 371-391.

Annex 1. Manual of R codes

Introduction

Domain estimators are implemented for at-risk-of-poverty rate, poverty gap, quintile share and Gini coefficient. For poverty rate, we have implemented HT-based equation (24), GREG estimator (26), which is assisted by a model provided by the user, and EBP(Y) estimator (16). Other indicators, such as the share of persons with low educational attainment, can be estimated with the estimators of poverty rate. Poverty gap, quintile share and Gini coefficient require special attention, as they cannot be estimated by methods designed for estimation of totals or shares. Their default estimators defined by equations (27), (30) and (33) are available.

Predictors of poverty gap (31), quintile share (34) and Gini coefficient (28) are implemented. However, it is preferable to apply the expansion technique (18) with log-transformation log(x+c+1) taking negative incomes into account as explained on p. 31. It is available for each predictor. Zero incomes are not processed separately in contrast with our simulation experiments. If the user has no unit-level population information about auxiliary variables, it is possible to use a frequency-calibrated predictor (19).

Composite estimators (Eqs. 20 and 21) are constructed from the default (direct) estimator and corresponding ordinary, expanded or frequency-calibrated predictor. The variance of the direct estimator is calculated by bootstrap.

Implementation

Our collection of R functions contains separate functions for default estimators (such as *direct_gini*) and predictors (e.g. *predictor_quintile_share*) in files gini.r, poverty_rate_estimator.r, poverty_gap_estimator.r and quintile_share.r, but the user does not have to call these directly (see next section for interface). Direct estimators and ordinary predictors are implemented by a call of function *domain_estimators* (in domain_estimators.r). Expanded predictors (Eq. 18) are implemented by a call of function *expanded_domain_predictors* (in domain_estimators.r) with the required predictor function as one of the arguments. The expanded predictions are calculated in

function *log_expanded_predictions* (expanded_predictions.r) A frequency-calibrated predictor is obtained by function *calibrated_predictors* (in calibrated_predictions.r), with predictor function as argument. A composite estimator is obtained by function *composite_estimators* (in composite_estimators.r), whose arguments include the direct estimator, the predictor, and the type of predictor (expanded, calibrated, or ordinary).

To support domain estimation, class Domain (domain.r) has methods finding all domains in a data set, and methods calculating domain indicators or domain sums, for example. The file estimated_distribution_function.r contains functions for calculating percentiles, among others.

Some special cases of data require somewhat arbitrary decisions. In the direct poverty gap estimator, all poor people of the sample are used if there are no poor in a domain. Similarly, the value of the poverty gap predictor is calculated from all predictions, if all predicted incomes in a domain exceed the poverty line. If a sample domain does not contain any observations, direct estimator is invalid, and the direct estimate is replaced by an estimate calculated with a predictor specified by the user.

Bootstrap samples are drawn by SRSWOR (R function *sample*) from a bootstrap population. The bootstrap population can be regarded as created by cloning each observation in the original sample with frequency equal to downwards rounded design weight. The bootstrap variance of a domain estimator is calculated as sample variance over bootstrap samples. The final composite weights are equal to the median over all domain-specific composite weights, irrespective of domain size.

Interface

The complexities of the implementation are hidden from an ordinary user. All the estimators of poverty indicators can be invoked through a single function *domain_estimate_data* (in interface.r). It creates a data set (R data frame) containing domain estimates for each domain.

The user has to fit a model to the sample and provide a function transforming the predictions to the original scale. Our R code assumes that the predicted values of a model can be obtained by calling generic R function *predict* with the model as the

first argument. This is possible with models fitted by *lm*, *glm*, *lme* and *nlme* (library nlme), but not necessarily with models of package lme4, for example.

Our R functions do not perform classification of variables. As an example, age classes must be created prior to domain estimation.

Poverty rate estimators are based on poverty indicators. They are first created by function *create_poverty_indicator* (in poverty_rate_estimator.r) which has the following arguments: sample, name of y variable, name of weights and the data set determining the poverty line (typically the sample). Then a logistic fixed-effects model is fitted by *glm* with option family=binomial or a logistic mixed model is fitted by *nlme*.

In the case of poverty gap, quintile share and Gini coefficient, a mixed model is usually fitted to log-transformed equivalized incomes by *lme*, for example. For log-transformation, the package includes functions *logp* and *expm*. *logp*(c) returns a function $f(x) = \log(x + c)$, and *expm*(c) returns its inverse function $f^{-1}(x) = \exp(x) - c$. If the model has been fitted to observations transformed by *logp*(c), then the corresponding back-transformation function is *expm*(c).

The estimators are specified by a list of names (argument estimator_descriptions of *domain_estimate_data*). The name of an estimator consists of the name of the poverty indicator and the type of the estimator. Names of the poverty indicators are ”poverty rate”, ”gini”, ”poverty gap” and ”quintile share”. Default estimators are identified by ”direct”, and predictors are identified by ”predictor”. Special cases of predictors are ”expanded” for predictors incorporating expanded predictions (18) and ”calibrated” for predictors based on the frequency-calibration (n-calibration) technique (19). In the case of poverty rate, it is also possible to use “greg” for GREG or MLGREG estimation and “ebp” for EBP estimation. Examples of estimator names are ”direct poverty rate”, ”greg poverty rate”, ”ebp poverty rate”, ”expanded gini predictor”, ”poverty gap predictor” and ”calibrated quintile share predictor”. The name of a composite estimator consists of the name of the unbiased component and the name of

the predictor, separated by a ”+”. An example is ”direct quintile share + expanded quintile share predictor”.

The domains are defined by a cross-tabulation of variables. A list of variable names is provided (argument domain_variables). The list can contain a single name, if the values of a variable are interpreted as domains. The domain variables must be present both in sample and in population.

If frequency-calibrated predictors are used, the population data set is still unit-level but one observation in each domain is chosen to contain the domain sums of those auxiliary variables that are used in calibration; the other observations of such auxiliary variables are zeroes.

The arguments of the function *domain_estimate_data* are as follows.

Argument	**Description**
estimator_descriptions	List of names of estimators
sample	Sample data (data frame)
population	Population data (data frame)
y	Name of the y variable
model	Model object. Function calls predict(model, newdata=population) and predict(model, newdata=sample) must work
back_transformation	Function back-transforming the predictions
x_list	List of names of quantitative x-variables used in n-calibration (or empty list)
xq_list	List of names of qualitative x-variables used in n-calibration (or empty list)
unknown	List of names of x-variables whose domain totals are estimated by GREG in n-calibration (or empty list)
domain_variables	List of names of variables determining the domains (crosstabulation)
weight	Name of the design weight variable in sample
reference_set	Data set determining the poverty line, typically sample
percentages	Vector of percentage points used in the expansion of predictions (Eq. 18); default is 1:99
missing_handler	Name of the type of predictor used to replace invalid direct estimates; examples: “expanded predictor”, “calibrated predictor”. Such a predictor is created for each poverty indicator.

Next excerpt of code is an example of poverty rate estimation by EBP based on a logistic mixed model (variable y is the equivalized income, w is the weight variable, x is an auxiliary variable and domain is the domain variable; pop is the population data set; *invlogit* is the function $\exp(x)/(1+\exp(x))$ provided in the package). Note that the poverty indicator has to be created and added to the sample, and its name "ind" is used as argument y in the call of function *domain_estimate_data*.

```
sample = data.frame(y,w,x,domain)
ind = create_poverty_indicator(sample,"y","w",sample)
data[["ind"]] = ind
model <- nlme(ind ~ invlogit(fix+ran), fixed=fix~x,
random=ran~1|domain, start=c(0,0))

back_transformation=identity

estimator_data <- domain_estimate_data(list("ebp poverty rate"),
sample=sample, population=pop, y="ind", model, back_transformation,
domain_variables=list("domain"), weight="w", reference_set=sample)
```

In the following example the resulting data set contains domain estimates by direct quintile share estimator, expanded quintile share predictor and their composite. The example presumes variables y, x and domain and data sets sample and pop as in previous example.

```
logy <- logp(1)(y)
model <- lme(logy ~ x, random=~1|domain)
back_transformation=expm(1)

estimator_data <- domain_estimate_data(list("direct quintile share",
"expanded quintile share predictor", "direct quintile share +
expanded quintile share predictor"), sample=sample, population=pop,
y="y", model, back_transformation, domain_variables=list("domain"),
weight="w", reference_set=sample, missing_handler = "expanded
predictor")
```

More detailed description of R codes is in Veijanen and Lehtonen (2011).

Annex 2. AMELI WP 2 Estimation: SUMMARY of SAE methods
Estimators of poverty indicators examined in simulations with register data (Finland)

Estimator	Description	Model	Aux. info
ESTIMATORS BASED ON INDICATOR VARIABLES			
At-risk-of poverty rate			
Design-based estimators			
1. DEFAULT	Design-based direct Horvitz-Thompson estimator	None	None
2. HT-CDF	Direct Horvitz-Thompson estimator based on cumulative distribution function	None	None
3. GREG	Generalized regression (GREG) estimator	Linear fixed-effects model	Area-level
4. MC	Model calibration estimator	Logistic fixed-effects model	Unit-level
5. LGREG	Logistic GREG estimator	Logistic fixed-effects model	Unit-level
6. MLGREG	Mixed-model assisted logistic GREG estimator	Logistic mixed model	Unit-level
Model-based estimators			
7. LSYN	Logistic synthetic estimator	Logistic fixed-effects model	Unit-level
8. EBP	Empirical best predictor incorporating predictions	Logistic mixed model	Unit-level
9. EBP(Y)	Empirical best predictor incorporating observations and predictions (EBLUP type)	Logistic mixed model	Unit-level
ESTIMATORS BASED ON MEDIANS AND QUANTILES			
Relative median at-risk-of poverty gap			
Design-based estimators			
10. DEFAULT	Design-based direct estimator	None	None
Model-based estimators			
11. SYN	Synthetic estimator based on mixed model predictions in population domain	Linear mixed model	Unit-level
12. SYN-EP	Synthetic estimator based on expanded (transformed) mixed model predictions in population domain	Linear mixed model	Unit-level
13. SYN-LOG	Synthetic estimator based on log-expanded (transformed) mixed model predictions in population domain	Linear mixed model	Unit-level
14. SYN-SIM	Synthetic simulation-based estimator (Molina and Rao 2010)	Linear mixed model	Unit-level
15. SYN-CAL	Calibrated synthetic estimator based on log-expanded (transformed) mixed model predictions in population domain	Linear mixed model	Area-level
Composite estimators			
16. COMP	Composite with DEFAULT and SYN-EP, MSE with nonparametric bootstrap	Linear mixed model	Unit-level
17. COMP-PB	Composite with DEFAULT and SYN-EP, MSE with parametric bootstrap	Linear mixed model	Unit-level
18. COMP-L	Composite with DEFAULT and SYN-LOG	Linear mixed model	Unit-level
19. COMP-C	Composite with DEFAULT and SYN-CAL	Linear mixed model	Area-level
Quintile share ratio (S20/S80 ratio)			
Design-based estimators			
20. DEFAULT	Design-based direct estimator	None	None
Model-based estimators			
21. SYN	Synthetic estimator based on mixed model predictions in population domain	Linear mixed model	Unit-level
22. SYN-EP	Synthetic estimator based on expanded (transformed) mixed model predictions in population domain	Linear mixed model	Unit-level
23. SYN-LOG	Synthetic estimator based on log-expanded (transformed) mixed model predictions in population domain	Linear mixed model	Unit-level
24. SYN-CAL	Calibrated synthetic estimator based on log-expanded (transformed) mixed model predictions in population domain	Linear mixed model	Area-level
Composite estimators			
25. COMP	Composite with DEFAULT and SYN-EP	Linear mixed model	Unit-level
26. COMP-L	Composite with DEFAULT and SYN-LOG	Linear mixed model	Unit-level
27. COMP-C	Composite with DEFAULT and SYN-CAL	Linear mixed model	Area-level

The Gini coefficient			
Design-based estimators			
28. DEFAULT	Design-based direct estimator	None	None
Model-based estimators			
29. SYN	Synthetic estimator based on mixed model predictions in population domain	Linear mixed model	Unit-level
30. SYN-EP	Synthetic estimator based on expanded (transformed) mixed model predictions in population domain	Linear mixed model	Unit-level
Composite estimators			
31. COMP	Composite with DEFAULT and SYN-EP	Linear mixed model	Unit-level
Estimators proposed for further investigation are in red.			

ANNEX 3 Technical SUMMARY of selected estimator types

AT - RISK - OF POVERTY RATE

Poverty indicator $v_k = I\{y_k \le \hat{t}_{HT}\}$ equals 1 for persons with income smaller than the estimated at-risk-of-poverty threshold $\hat{t}_{HT} = 0.6\hat{M}$ and 0 for others, where $\hat{M}$ refers to median estimate

DEFAULT (HT) estimator (1) $\hat{r}_{d;HT} = \sum_{k\in s_d} a_k v_k / N_d, \quad \hat{r}_{d;HT} = \sum_{k\in s_d} a_k v_k / \hat{N}_d, \quad d = 1,\ldots,D$

GREG, LGREG and MLGREG estimators (3, 5, 6) $\hat{r}_{d;GREG} = \hat{f}_{d;GREG} / N_d, \qquad \hat{r}_{d;GREG} = \hat{f}_{d;GREG} / \hat{N}_d$

where $\hat{f}_{d;GREG} = \sum_{k\in U_d} \hat{v}_k + \sum_{k\in s_d} a_k \hat{e}_k$

and $a_k = 1/\pi_k$, $\hat{e}_k = v_k - \hat{v}_k$, N_d is size of population domain and $\hat{N}_d = \sum_{k\in s_d} a_k$

Model calibration MC estimator (4) $\hat{r}_{d;MC} = \sum_{k\in s_d} w_{rk} v_k / N_d, \ \hat{r}_{d;MC} = \sum_{k\in s_d} w_{rk} v_k / \hat{N}_d$

where $\sum_{k\in s_r} w_{rk}\mathbf{z}_k = \sum_{k\in U_r} \mathbf{z}_k$ and $\mathbf{z}_k = (1,\hat{v}_k)'$

LSYN and EBP estimators (7, 8) $\hat{r}_{d;SYN} = \sum_{k\in U_d} \hat{v}_k / N_d, \quad \hat{r}_{d;SYN} = \sum_{k\in U_d} \hat{v}_k / \hat{N}_d$

EBP(Y) estimator (9) $\hat{r}_{d;EBP(Y)} = (\sum_{k\in U_d - s_d} \hat{v}_k + \sum_{k\in s_d} v_k) / N_d$

Predictions for GREG $\hat{v}_k = \mathbf{x}_k'\hat{\boldsymbol{\beta}}, \quad k \in U$

Predctions for LGREG, LSYN and MC $\hat{v}_k = \dfrac{\exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}})}$

Predictions for MLGREG, EBP and EBP(Y) $\hat{v}_k = \dfrac{\exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_{0r})}{1+\exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_{0r})}$

POVERTY GAP

Value y_k of study variable y refers to equalized income (transformed $\log(y_k + 1)$ was used in model fitting)

DEFAULT estimator (10) $\hat{g}_d = \dfrac{\hat{t} - Md\{y_k; y_k \le \hat{t}; k \in s_d\}}{\hat{t}}$

SYN type estimators (11-15) $\hat{g}_{d;SYN} = \dfrac{\hat{t} - Md\{\hat{y}_k; \hat{y}_k \le \hat{t}; k \in U_d\}}{\hat{t}}$

Predictions for SYN type estimators $\hat{y}_k = \exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_{0r}) - 1$

Composite type estimators (16-19) $\hat{g}_{d;COMP} = \hat{\lambda}_d \hat{g}_d + (1-\hat{\lambda}_d)\hat{g}_{d,SYN}$

where $\hat{\lambda}_d$ is average of $\dfrac{M\hat{S}E(\hat{g}_{d,SYN})}{M\hat{S}E(\hat{g}_{d,SYN}) + M\hat{S}E(\hat{g}_d)}$ over a domain size class

QUINTILE SHARE RATIO (S20/S80 ratio)

DEFAULT estimator (20) $\hat{q}_d = \hat{S}20_d / \hat{S}80_d$

where $\hat{S}20_d = \dfrac{\sum_{k\in q_{d,20}} a_k y_k}{\sum_{k\in q_{d,20}} a_k}$ and $\hat{S}80_d = \dfrac{\sum_{k\in q_{d,80}} a_k y_k}{\sum_{k\in q_{d,80}} a_k}$

and $q_{d,20}$ (first quintile) is the set of poorest people in domain d whose sum of weights is just below or at 20% of the total sum of weights ($q_{d,80}$ similarly)

SYN type estimators (21-24) $\hat{q}_{d,SYN} = \hat{S}20_{d,SYN} / \hat{S}80_{d,SYN}$

where $\hat{S}20_{d;SYN} = \dfrac{\sum_{k\in q_{SYN;d,20}} \hat{y}_k}{\sum_{k\in U_d} I\{k \in q_{SYN;d,20}\}}$ and $\hat{S}80_{d;SYN} = \dfrac{\sum_{k\in q_{SYN;d,80}} \hat{y}_k}{\sum_{k\in U_d} I\{k \in q_{SYN;d,80}\}}$

and $q_{SYN;d,20}$ denotes the fifth quintile defined in population domain as if the weights were constant

Predictions for SYN type estimators: $\hat{y}_k = \exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_{0r}) - 1, \quad k \in U$

Composite type estimators (25-27) $\hat{q}_{d;COMP} = \hat{\lambda}_d \hat{q}_d + (1-\hat{\lambda}_d)\hat{q}_{d,SYN}$

where $\hat{\lambda}_d$ is average of $\dfrac{M\hat{S}E(\hat{q}_{d,SYN})}{M\hat{S}E(\hat{q}_{d,SYN}) + M\hat{S}E(\hat{q}_d)}$ over a domain size class