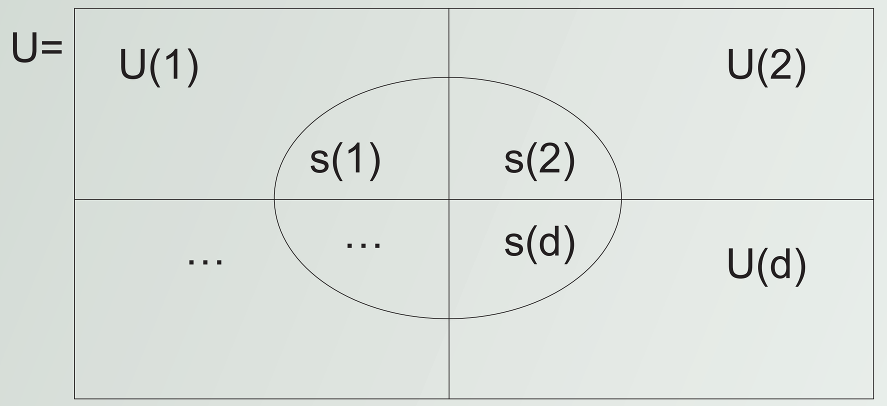
The Role of Models in Model-Assisted and Model-Dependent Estimation for Small Areas

Risto Lehtonen, University of Helsinki
Mikko Myrskylä, University of Pennsylvania
Carl-Erik Särndal, Université de Montréal
Ari Veijanen, Statistics Finland

Background

- There is an increasing need for Small Area Estimation (SAE), that is estimation of statistics for regional and other domains
- Some recent Small Area Estimation projects:
  - SAIPE, U.S. Census Bureau's model-based Small Area Income and Poverty Estimation project
  - EURAREA Project, Adaptation of model-dependent small area estimation methods into the European context
- Auxiliary information and statistical models have a crucial role in small area estimation
- This poster studies the accuracy of two conventional small area estimators and a new, hybrid estimator under various model formulations

SAE task: estimate Y(d)



- Population U, sample s, weights w
- Target: Y(d) = total of y in domain U(d)
- Simple but inefficient Horvitz-Thompson estimator:

$$\hat{Y}(d) = \sum_{s(d)} w_i y_i$$

Efficient estimators

The most commmonly used SAE estimators GREG (1) and EBLUP (2) utilize auxiliary information x, which is used to predict study variable y:

$$\hat{y}_i = f(x_i; \hat{\beta}).$$

1) Model-assisted Generalized regression estimator GREG (Does take into account sampling weights):

$$\hat{Y}_G(d) = \sum_{U(d)} \hat{y}_i + \sum_{s(d)} w_i (y_i - \hat{y}_i)$$

2) Model-dependent EBLUP (Relies on the model, ignores sampling design):

$$\hat{Y}_E(d) = \sum_{s(d)} y_i + \sum_{U(d)-s(d)} \hat{y}_i$$

3) Weighted EBLUP: Estimator formulation like EBLUP, but sampling design is taken into account by using weights in model estimation

Known properties of estimators

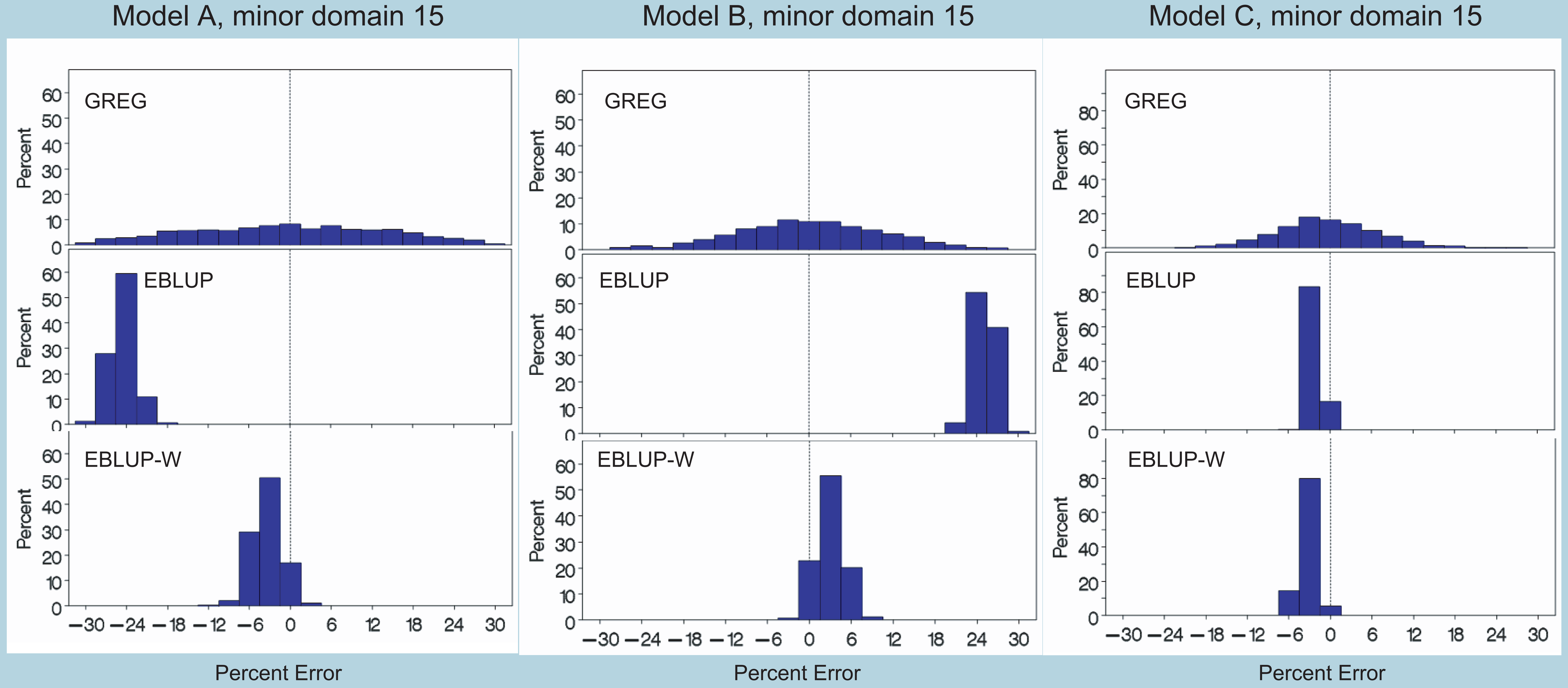
	Model-assisted GREG	Model-dependent EBLUP	Weighted EBLUP*
Bias	Approximately unbiased	Bias depends on the model	?
Variance	Variance usually large for small domains	Variance may be small even for small domains	?
Mean square error	MSE approximately equal to variance	MSE dominated by bias	?
Conf. intervals	Valid intervals can be constructed	Valid intervals not necessarily obtained	?

*Properties of Weighted EBLUP (EBLUP-W) are not known since this estimator has not been used before

Results of the Monte Carlo study

Estimator	Model	Average absolute relative bias (%)		Average relative root MSE (%)	
		Expected domain sample size		Expected domain sample size	
		Minor (20-69)	Major (120+)	Minor (20-69)	Major (120+)
GREG	A $y_i = \beta_0 + u_d + \varepsilon_i$	0.2	0.1	13.7	5.6
	B $y_i = \beta_0 + u_d + \beta_2 x_{2i} + \varepsilon_i$	0.2	0.1	11.6	4.8
	C $y_i = \beta_0 + u_d + \beta_1 x_{1i} + \varepsilon_i$	0.2	0.0	7.8	3.3
EBLUP	A $y_i = \beta_0 + u_d + \varepsilon_i$	22.9	21.7	22.9	21.8
	B $y_i = \beta_0 + u_d + \beta_2 x_{2i} + \varepsilon_i$	22.3	21.8	22.4	21.9
	C $y_i = \beta_0 + u_d + \beta_1 x_{1i} + \varepsilon_i$	1.8	0.7	2.8	2.2
Weighted EBLUP	A $y_i = \beta_0 + u_d + \varepsilon_i$	3.7	3.3	3.9	3.5
	B $y_i = \beta_0 + u_d + \beta_2 x_{2i} + \varepsilon_i$	3.7	3.2	3.9	3.3
	C $y_i = \beta_0 + u_d + \beta_1 x_{1i} + \varepsilon_i$	3.5	3.3	3.5	3.3

Monte Carlo error distribution of the estimators in one domain



Design of the Monte Carlo study

- This Monte Carlo study compares the accuracy of GREG, EBLUP, and EBLUP-W under various model formulations
- Population: N = 1,000,000, divided into 100 domains
- Samples: K = 1000 PPS samples of size n = 10,000
- Sampling weights vary between 54.6 and 596.5
- Study variable is generated as

$$y_i = 1 + 2x_{1i} + 1.5x_{2i} + u_d + \varepsilon_i$$

- For every sample k domain totals are estimated using GREG, EBLUP, EBLUP-W and three different models

Summary of results

- Model-assisted GREG
  - Approximately unbiased for all models
  - Variance large in small areas
  - Accurate if domain sample size is large
- Model-dependent EBLUP
  - Severy biased if model is not good
  - Variance small even if model is weak
  - Accurate if model is very good
- Weighted EBLUP
  - Bias relatively small for all models
  - Variance small even if model is weak
  - Relatively accurate even for weak models and small areas

Some relevant literature

- Estevao, V.M. and Särndal, C.-E. (1999) The use of auxiliary information in design-based estimation for domains. *Survey Methodology*, 25, 213–221.
- Fay, R.E. and Herriot , R.A. (1979) Estimation of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269–277.
- 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.
- Rao, J.N.K. (2003) *Small Area Estimation*. Hoboken: Wiley.
- Särndal, C.-E., Swensson, B. and Wretman, J.H. (1992) *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Penn Demography