CALIBRATION OF SMALL AREA ESTIMATES IN BUSINESS SURVEYS

Rodolphe Priam, Natalie Shlomo

Southampton Statistical Sciences Research Institute
University of Southampton
United Kingdom

SAE, August 2011

The BLUE-ETS Project is financed by the grant agreement no:
244767 under Theme 8 of the 7th Framework Programme (FP7) of the European Union, Socio-economic Sciences and Humanities.

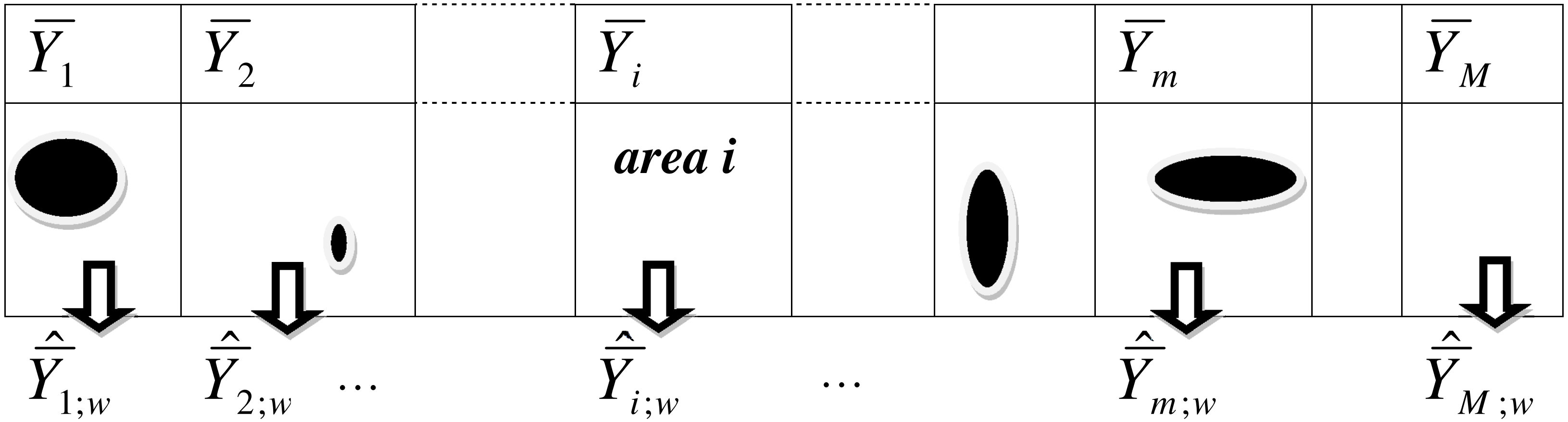
BUSINESS SURVEYS

- Statistical units are organisational entities in a country
- Interested in small area/domain estimates
- Business registers allow for unit level covariates
- Distributions are typically skewed with outliers
- Transformations, such as the log, to ensure normality assumptions

SMALL AREA ESTIMATION

- **Central problem** in many areas of social statistics. Recently used in business statistics.

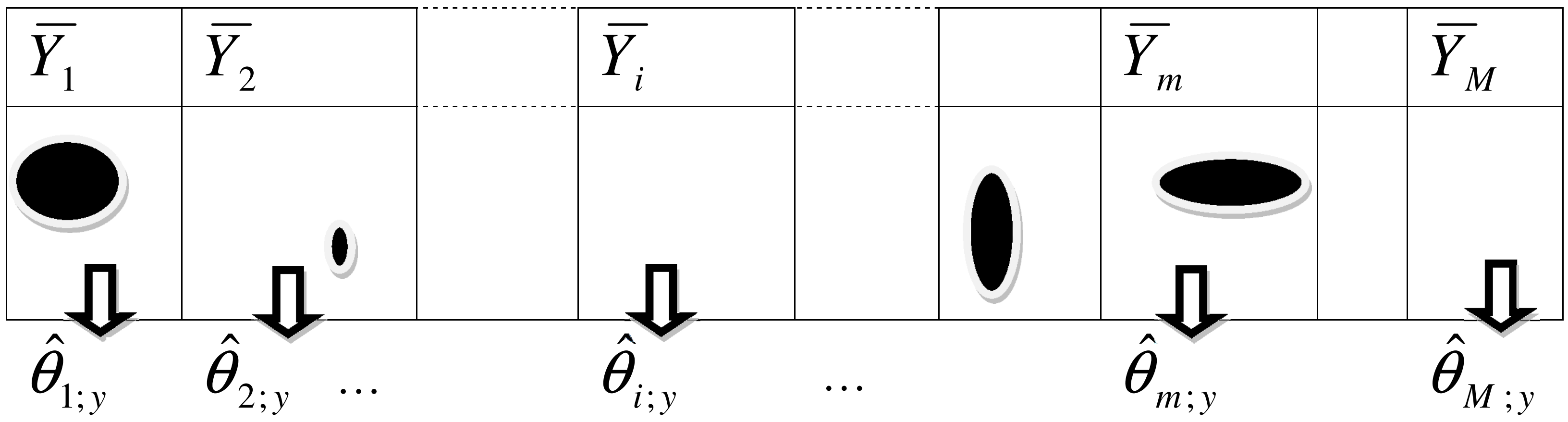
- Estimation of the mean in diverse domains



- True population mean $\bar{Y}_i$ and design-based estimate $\hat{\bar{Y}}_{i;w}$
- Estimated small area mean (EBLUP) $\hat{\theta}_{i;y}$ because of small n_i

SMALL AREA ESTIMATION AND BENCHMARKING

- Small area estimation of the total in the different domains



<u>Problem</u>: The total estimated by the model $\tilde{T}_y = \sum_i w_i \hat{\theta}_{i;y}$ should match the design based estimate of the population total $\hat{T}_y = \sum_i w_i \hat{\overline{Y}}_{i;w}$.

- Solution by benchmarking the estimates by appropriate method
- Consequence of more robust estimation to misspecifications of the model.

NESTED ERROR UNIT LEVEL MODEL

- The Battese, Harter and Fuller (1988) (BHF) model for small areas *i=1, …, M*:

$$\mathrm{Y}_i = \mathrm{X}_i \beta + 1_{N_i} u_i + \mathrm{e}_i$$

- The target parameter of interest is the area mean:

$$\bar{Y}_i = 1'_{N_i} Y_i / N_i$$

- The EBLUP for non-negligible sampling fractions:

$$\hat{\theta}^{f}_{i;y} = f_i \bar{y}_i + (1 - f_i)\left[\bar{X}'_{ic}\,\hat{\beta}_{GLS} + \hat{u}_i\right]$$

BENCHMARKING AT THE LINEAR SCALE (1/2)

- Existing methods considered (see for instance Wang & al. (2008))
  - The ratio method by multiplicative term: $\hat{\theta}_{i;y}^{RT} = \hat{T}_y \tilde{T}_y^{f\,-1} \hat{\theta}_{i;y}^{f}$
  - An additive term with variance weighting: $\hat{\theta}_{i;y}^{VAR} = \hat{\theta}_{i;y}^{f} + \dfrac{N_i\left(\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_i\right)}{\sum_{i=1}^{m} N_i^2\left(\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_i\right)}\left(\hat{T}_y - \tilde{T}_y^{f}\right)$
  - Pfeffermann and Barnard (1991): $\hat{\theta}_{i;y}^{PB} = f_i \bar{y}_i + (1 - f_i)\left[\bar{X}_{ic}' \hat{\beta}_{PB} + \hat{u}_i^{PB}\right]$

where $\hat{\eta}^{PB} = \hat{\eta} - CR'(r - R\hat{\eta}) / RCR'$, $\hat{\eta} = (\hat{\beta}_{GLS}', \hat{u}_1, \ldots, \hat{u}_M)'$, $r = \hat{T}_y - n\bar{y}$, $R\hat{\eta}^{PB} = r$,

$$R = \left(\sum_{i=1}^{M} N_i \bar{X}_i, N_1 - n_1, N_2 - n_2, \cdots, N_m - n_m, N_{m+1}, \cdots, N_M\right)$$

Ugarte & al. (2009) applied this constrained model for a business survey for several regions with variance calculations

BENCHMARKING AT THE LINEAR SCALE (2/2)

- We propose the method

  Augmentation of the unconstrained least-squares system by adding to the original GLS system one row and one column:

$$\begin{pmatrix} y_s \\ y_{+;a} \end{pmatrix} = \begin{pmatrix} X_s & w_a \\ X'_{+;a} & w_{+;a} \end{pmatrix} \beta_{PSW} + e_a = \begin{pmatrix} X_{s;a} \\ X'_{+;a} \end{pmatrix} \beta_{PSW} + e_a$$

where,

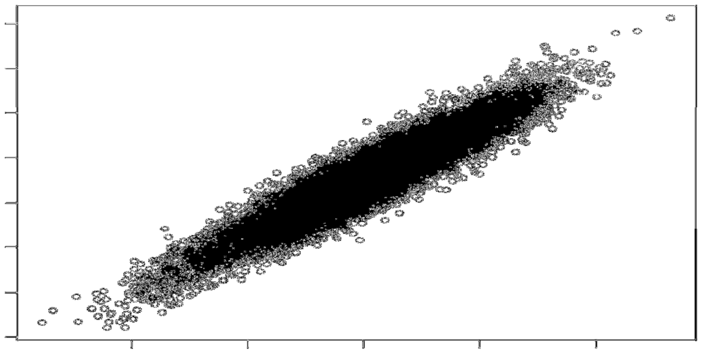
$$w_a = \left(w'_{1;a}, w'_{2;a}, \cdots, w'_{m;a}\right)' ;\ w_{i;a} = \left(N_i / n_i - 1\right) \times 1_{Ni} ;\ X'_{+;a} = \sum_{i=1}^{m} \left(N_i - n_i\right)\left\{- \overline{X}'_{ic;a} + (2\hat{\gamma}_i - 1)\overline{x}'_{i;a}\right\} ;$$

$$y_{+;a} = \sum_{i=1}^{m} \left((2\hat{\gamma}_i - 1)\left(N_i - n_i\right) + n_i \left(1 - N / n\right)\right)\overline{y}_i ;\ w_{+;a} = 2\sum_{i=1}^{m} (\hat{\gamma}_i - 1)\left(N_i - n_i\right)^2 / n_i .$$

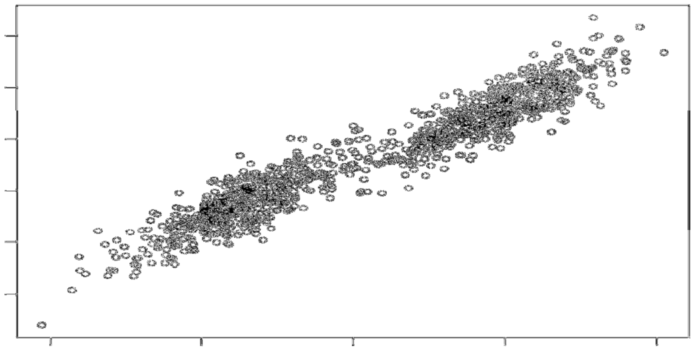
- The benchmarking equation is obtained by orthogonality of the residual to the new added column

SIMULATION FOR LINEAR CASE

- Nested error unit level regression model
- *B=1000* populations generated
- M = 30 areas (no empty areas)
- $f_i \approx 4\%$
- $\sigma_u = 0.1$, $\sigma_e = 0.3$, and $\beta = (2, 0.25)^T$
- $x_{ij} \sim N(m_i, s_i)$; $m_i \sim N(10,3)$; $s_i = 2$



ONE POPULATION GENERATED



TWO AREAS IN THE POPULATION

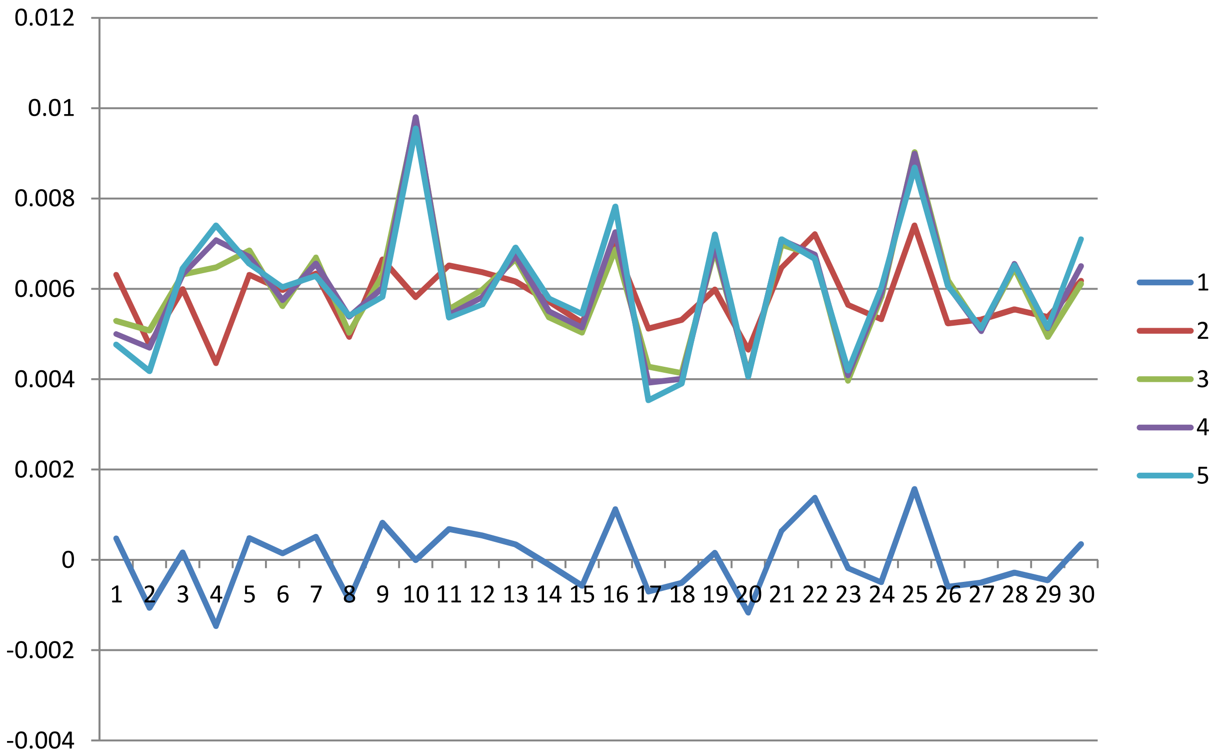
SIMULATION RESULT FOR LINEAR CASE (1/2)

1	$\hat{\theta}^{f}_{i;y}$	EBLUP
2	$\hat{\theta}^{RT}_{i;y}$	Ratio Benchmark
3	$\hat{\theta}^{VAR}_{i;y}$	Variance Weighted Benchmark
4	$\hat{\theta}^{PB}_{i;y}$	Pfeffermann and Barnard Benchmark
5	$\hat{\theta}^{PSW}_{i;y}$	Proposed Method Benchmark

	1	2	3	4	5
	$\hat{\theta}^{f}_{i;y}$	$\hat{\theta}^{RT}_{i;y}$	$\hat{\theta}^{VAR}_{i;y}$	$\hat{\theta}^{PB}_{i;y}$	$\hat{\theta}^{PSW}_{i;y}$
BIASREL	0.06%	0.58%	0.60%	0.60%	0.60%
AARB	0.04%	0.60%	0.62%	0.62%	0.62%
ARMSE	1.31%	1.45%	1.46%	1.46%	1.47%
DIFFTOT	4.0×10^{2}	0.000	0.000	0.000	0.000

SIMULATION RESULT FOR LINEAR CASE (2/2)

1	$\hat{\theta}_{i;y}^{f}$	EBLUP
2	$\hat{\theta}_{i;y}^{RT}$	Ratio Benchmark
3	$\hat{\theta}_{i;y}^{VAR}$	Variance Weighted Benchmark
4	$\hat{\theta}_{i;y}^{PB}$	Pfeffermann and Barnard Benchmark
5	$\hat{\theta}_{i;y}^{PSW}$	Proposed Method Benchmark



LOG TRANSFORMATION FOR SKEWED VARIABLE

- In BHF model,

$$y_{ij} = x_{ij}\beta + u_i + e_i$$

- In business surveys, distributions are skewed
  - Log normal transformation

$$z_{ij} = \exp(x_{ij}\beta + u_i + e_i)$$

  - New formulation of the predictors

BACK-TRANSFORMATION WITH BIAS CORRECTION

- Formulation of a nearly unbiased estimator is:

$$\hat{\theta}_{i;z}^{f,sum} = f_i \bar{z}_i + (1 - f_i) \sum_{j \in U_i \setminus s_i} \exp(\hat{y}_{ij} + \hat{\alpha}_i) \quad (1)$$

The bias correction is $\hat{\alpha}_i$ and can be defined at the unit level or area level (see Chambers, Dorfman (2003) and Molina (2009))

- Other formulation from Kurnia, Notodiputro, Chambers (2009):

$$\hat{\theta}_{i;z}^{*,\exp} = \exp(\hat{\theta}_{i;y}^{*} + \tilde{\alpha}_i) \quad (2)$$

  - The bias correction is the modified term at the area level $\tilde{\alpha}_i$
  - We propose the corrective term $\tilde{\alpha}_{i2}$ and compare to $\tilde{\alpha}_{i1}$

$$(a)\ \tilde{\alpha}_{i1} = \hat{\alpha}_i$$

$$(b)\ \tilde{\alpha}_{i2} = \hat{\alpha}_i + \frac{1}{2}\hat{\beta}^T \hat{\Sigma}_i \hat{\beta}$$

where $\hat{\Sigma}_i$ is the covariance matrix of the covariates.

BACK-TRANSFORMATION WITH BIAS CORRECTION

- Approaches under model (1)
  - Chambers, Dorfman (2003) introduce several estimators: the rast predictor and smearing predictor
  - Fabrizi, Ferrante, Pacei (2007) compare estimators to a naïve predictor without a bias correction. The twiced smeared estimator performed best in simulation
  - Chandra, Chambers (2011) discuss calibration after a log-transformation

BENCHMARKING AFTER BACK-TRANSFORMATION

Compare benchmarking at different stages with back transformation and bias correction by: (a) $\hat{\alpha}_i = \left(\hat{\sigma}_u^2 + \hat{\sigma}_e^2\right)/2$ or (b) $\tilde{\alpha}_{i2} = \hat{\alpha}_i + \hat{\beta}'\hat{\Sigma}_i\hat{\beta}/2$

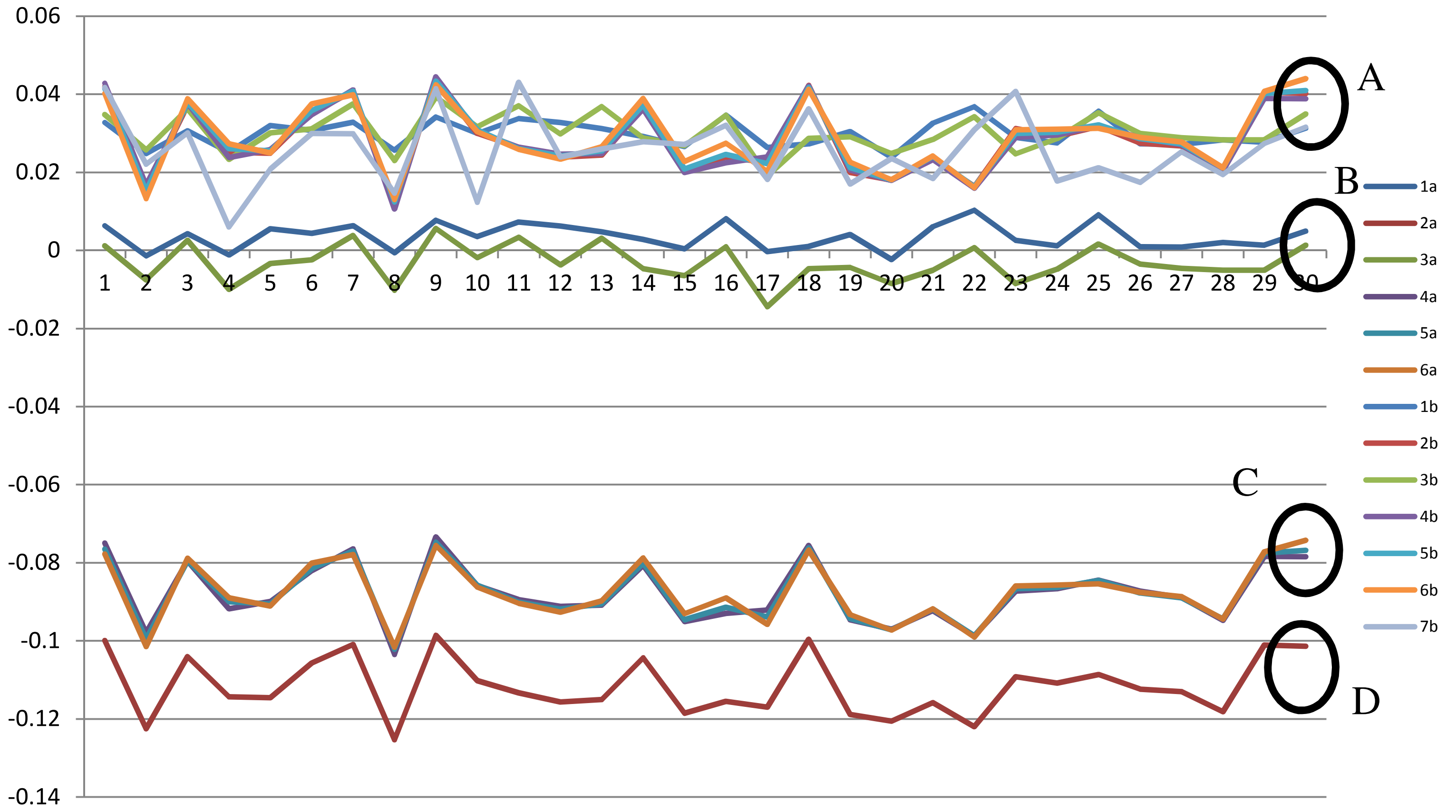
- Ratio method under different scenarios
  - No benchmark at log scale, back-transformed method (2), bias correction (a) $\hat{\theta}_{i;z}^{f,RT}$
  - Benchmark at log scale, back-transformed method (2), bias correction (a) $\hat{\theta}_{i;z}^{VAR,RT}$ $\hat{\theta}_{i;z}^{PB,RT}$ $\hat{\theta}_{i;z}^{PSW,RT}$
  - No benchmark at log scale, back-transformed method (1), bias correction (a) $\hat{\theta}_{i;z}^{f,sum,RT}$
  - No benchmark at log scale, back- transformed method (2), bias correction (b) $\hat{\theta}_{i;z}^{f2,RT}$
- A maximization of the log-likelihood of the BHF model under constraints, back transformed method (2) and bias correction (b) $\hat{\theta}_{i;z}^{MLC}$

SIMULATION RESULT FOR NON-LINEAR CASE (1/2)

- No benchmark at log scale, back-transformed method (2) , ,bias correction (a) , ratio adjusted $\hat{\theta}_{i;z}^{f,RT}$
- Benchmark at log scale, back- transformed method (2) , bias correction (a), ratio adjusted $\hat{\theta}_{i;z}^{VAR,RT}$ $\hat{\theta}_{i;z}^{PB,RT}$ $\hat{\theta}_{i;z}^{PSW,RT}$
- No benchmark at log scale, back- transformed method (1) , bias correction (a) , ratio adjusted $\hat{\theta}_{i;z}^{f,sum,RT}$
- No benchmark at log scale, back- transformed method (2) , bias correction (b), ratio adjusted $\hat{\theta}_{i;z}^{f2,RT}$
- MLC adjustment, back- transformed method (2) , bias correction (b) $\hat{\theta}_{i;z}^{MLC}$

	NOT BENCHMARKED						BENCHMARKED						
	1a	2a	3a	4a	5a	6a	1b	2b	3b	4b	5b	6b	7b
	$\hat{\theta}_{i;z}^{f,sum}$	$\hat{\theta}_{i;z}^{f}$	$\hat{\theta}_{i;z}^{f2}$	$\hat{\theta}_{i;z}^{VAR}$	$\hat{\theta}_{i;z}^{PB}$	$\hat{\theta}_{i;z}^{PSW}$	$\hat{\theta}_{i;z}^{f,sum,RT}$	$\hat{\theta}_{i;z}^{f,RT}$	$\hat{\theta}_{i;z}^{f2,RT}$	$\hat{\theta}_{i;z}^{VAR,RT}$	$\hat{\theta}_{i;z}^{PB,RT}$	$\hat{\theta}_{i;z}^{PSW,RT}$	$\hat{\theta}_{i;z}^{MLC}$
BIASREL	0.39%	11.16%	0.47%	8.77%	8.77%	8.75%	2.99%	2.84%	3.03%	2.83%	2.87%	2.90%	2.58%
AARB	0.66%	10.89%	0.28%	8.50%	8.49%	8.49%	3.30%	3.15%	3.34%	3.15%	3.18%	3.20%	2.89%
ARMSE	5.81%	12.05%	5.75%	10.01%	10.01%	10.02%	6.87%	6.84%	6.90%	6.84%	6.86%	6.90%	6.69%
DIFFTOT	$5.6x10^4$	$3.0x10^5$	$7.1x10^4$	$2.5x10^5$	$2.5x10^5$	$2.5x10^5$	0.00	0.00	0.00	0.00	0.00	0.00	0.00

SIMULATION RESULT FOR NON-LINEAR CASE (2/2)



Group A: All benchmark estimates to original scale using the Ratio Method or the MLC method ('1b' – '7b')
Group B: No benchmark, back- transformed method (1) and bias correction (a) ('1a') and back- transformed method (2) and bias correction (b) ('3a')
Group C: Benchmark at log-scale and no benchmark to original scale, back- transformed method (2) and bias correction (a) ('4a', '5a', '6a')
Group D: No benchmark, back-transformed method (2) and bias correction (a) ('2a')

CONCLUSION

- We have used the nested error unit level regression model
- Benchmarking methods for the linear case perform similarly
- Benchmarking methods for non-linear case differ depending on back-transformation and stage of benchmarking
- Ratio adjustment to benchmarked log-scale and back transformation provide comparable results to the case when log-scale is not benchmarked
- Future research:
  - ➢ Performance under more realistic populations, empty areas
  - ➢ Comparison with alternative methods, for example robust methods of small area models
  - ➢ Inclusion of survey weights, variance estimates

Thanks for your attention