

An Empirical Comparison of EBLUP Estimation and Model Based Direct Estimation for Small Areas

Hukum Chandra and Ray Chambers

University of Southampton, United Kingdom
and
University of Wollongong, Australia

Overview

- Linear mixed models
- Weights based on linear mixed models
- The EBLUP and its MSE estimator
- The MBD estimator
- MSE estimation for the MBD estimator
- Empirical results
- Conclusions

Background

- Small area estimation (SAE) is now very common in survey sampling and several methods for the estimation have been proposed in the literature (Rao, 2003)
- Unit level random effect models are often used in SAE
- The Empirical Best Linear Unbiased Prediction (EBLUP) approach (Prasad and Rao, 1990) is the most popular model based techniques for the SAE under such models
- These approaches typically do not use the unit level survey weights in their estimators

Background

- As result, simplicity of using linearly weighted estimators are lost
- The Model-Based Direct (MBD) approach (Chambers and Chandra, 2006) overcomes these limitations
- Uses calibrated sample weights derived via population level version of linear mixed model in SAE
- MBD Estimator: Defined as linearly weighted estimator

Linear Mixed Model

- **The most commonly used class of models in small area inference**

$$Y_i = X_i\beta + Z_iu_i + e_i; \quad i = 1, \dots, m$$

- Y_i is the $N_i \times 1$ vector of values of variable of interest Y in the small area i
- X_i is a $N_i \times p$ matrix of known auxiliary variables
- N_i is the number of the population units in the area i
- m is the number of small areas
- β is $p \times 1$ vector of fixed effects

- Z_i is a $N_i \times q$ matrix of known covariates
- u_i is the random area effect associated with the small area i
- e_i is the $N_i \times 1$ vector of random errors
- u_i and e_i are independent random vectors, both with zero mean vectors and with $Var(u_i) = \Sigma(\theta)$, $Var(e_i) = \sigma_e^2 I_{N_i}$ so that

$$Var(Y_i) = V_i = \sigma_e^2 I_{N_i} + Z_i \Sigma(\theta) Z_i'$$

- Aggregating **m-small area models**, lead to **population level linear mixed model (with block diagonal structure)**

$$Y = X\beta + Zu + e$$

with $V = Var(Y) = \text{block diagonal}(V_i)$

Mixed Model Weighting

- Estimate variance components θ and σ_e^2 from sample data, leading to $\hat{\theta}$ and $\hat{\sigma}_e^2 \Rightarrow \hat{V}_i = \hat{\sigma}_e^2 I_i + Z_i \Sigma(\hat{\theta}) Z_i' \Rightarrow \hat{V} = \text{block diagonal}(\hat{V}_i)$
- Use appropriate sample/Non-sample decompositions
- Under the population level mixed model, **EBLUP weights** (Royall, 1976)

$$W_{EBLUP} = 1_s + H'_{EBLUP} (X'1_N - X'_s 1_s) + (I_s - H'_{EBLUP} X'_s) \hat{V}_{ss}^{-1} \hat{V}_{sr} 1_r$$

$$H_{EBLUP} = \left(X'_s \hat{V}_{ss}^{-1} X_s \right)^{-1} X'_s \hat{V}_{ss}^{-1}$$

- 1_N , 1_n and 1_r are vectors of 1's of order N, n and (N-n) and I_s identity matrix of order n

The Industry Standard

EBLUP of the area i mean $\bar{Y}_i = N_i^{-1} \sum_{U_i} y_j$

$$\hat{Y}_i^{EBLUP} = f_i \bar{Y}_{is} + (1 - f_i) [\bar{X}'_{ir} \hat{\beta} + \bar{Z}'_{ir} \hat{\Sigma} Z'_{is} \hat{V}_{iss}^{-1} (Y_{is} - X_{is} \hat{\beta})]$$

MSE usually estimated via the Prasad-Rao estimator

$$mse(\hat{Y}_i^{EBLUP}) = (1 - f_i)^2 \left[g_{1i}(\hat{\theta}) + g_{2i}(\hat{\theta}) + 2g_{3i}(\hat{\theta}) \right] + N_i^{-1} (1 - f_i) \hat{\sigma}_e^2$$

where $\hat{\theta}$ and $\hat{\sigma}_e^2$ are the estimates of the variance components and g_{1i} , g_{2i} and g_{3i} are rather complicated functions

An Alternative

Model-Based Direct Estimator $\hat{Y}_{i,MBD} = \sum_{j \in s_i} w_j y_j / \sum_{j \in s_i} w_j$

- w_j are the EBLUP (mixed model) weights derived earlier
- EBLUP and MBD are **not** the same at small area level. However, both methods of small area estimation lead to the same overall population estimate

$$\sum_{i=1}^m N_i \hat{Y}_i^{EBLUP} = \sum_{i=1}^m \hat{N}_i \hat{Y}_i^{MBD} = \sum_{j \in s} w_j y_j$$

where $\hat{N}_i = \sum_{s_i} w_j$ and weights are the EBLUP weights w_j

MSE Estimation for the MBD Estimator

Adapt standard robust methods of MSE estimation for population totals and means

At population level $v(\hat{Y}) = \sum_{j \in s} w_j^2 (y_j - \hat{y}_j)^2 + \text{lower order terms}$

- An approximately unbiased estimate of $Var(y_j)$ is given as squared residual $(y_j - x'_j \hat{\beta})$ (Royall and Cumberland, 1978)
- **Estimate of MSE of MBD estimator** for the i^{th} small area mean (Chambers and Chandra, 2006)

$$mse(\hat{Y}_i^{MBD}) = v(\hat{Y}_i^{MBD}) + \left[\hat{bias}(\hat{Y}_i^{MBD}) \right]^2$$

Estimate of Prediction variance $v(\hat{Y}_i^{MBD}) = \sum_{j \in s_i} \lambda_j (y_j - x'_j \hat{\beta})^2$

$$\lambda_j = N_i^{-2} \left(a_j^2 + \frac{(N_i - n_i)}{(n_i - 1)} \right) \quad \text{and} \quad a_j = \left(N_i w_j - \sum_{s_i} w_j \right) / \left(\sum_{s_i} w_j \right)$$

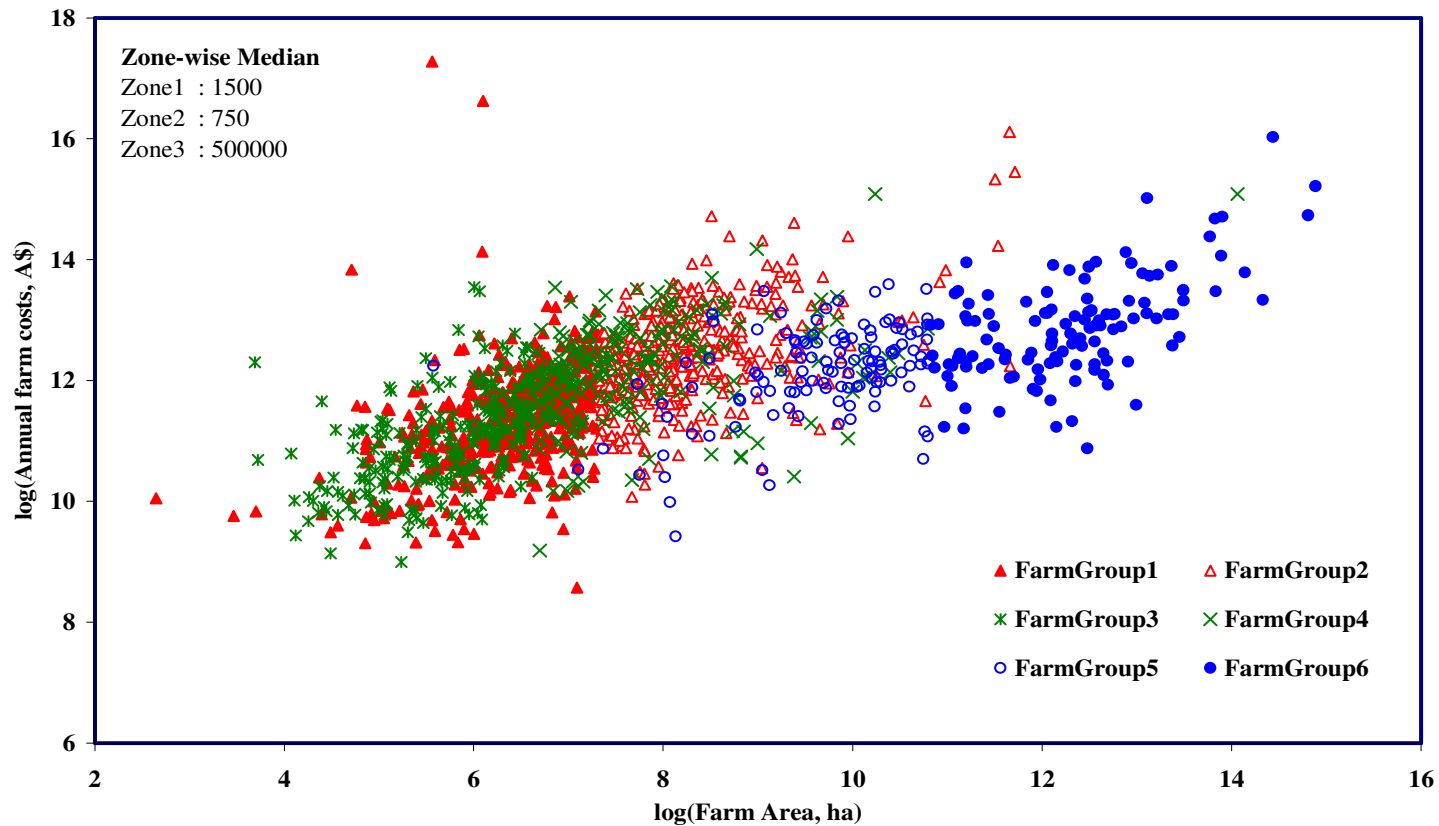
Estimate of bias $\hat{bias}(\hat{Y}_i^{MBD}) = (\hat{X}_i^{MBD} - \bar{X}_i)' \hat{\beta}$

- Bias correction due to using the above estimator on small area level
- \hat{X}_i^{MBD} is the weighted average of the sample x_j in the area i
- \bar{X}_i is the population mean of x_j 's in the small area i

Empirical Evaluation

- Australian broadacre farms survey data on 1652 sample farms
- Generated a target population of 81982 farms by sampling with replacement from them with probabilities proportional to their sample weights
- 29 different Australian broadacre agricultural **regions** was considered as 29 **small areas** of interest
- Sample size within small areas varied from 6 to 117
- Regions are grouped into zones (Pastoral, Mixed Farming, Coastal)
- **Six post strata**: splitting each zone into small farms (farm area < than zone median) and large farms (farm area > than or equal to zone median)

Average Farm Cost vs Average Farm Area in 6 Post Strata



Variables (Y) : Annual Farm Costs (A\$)
Auxiliary (X) : Farm Area

Two X Specifications

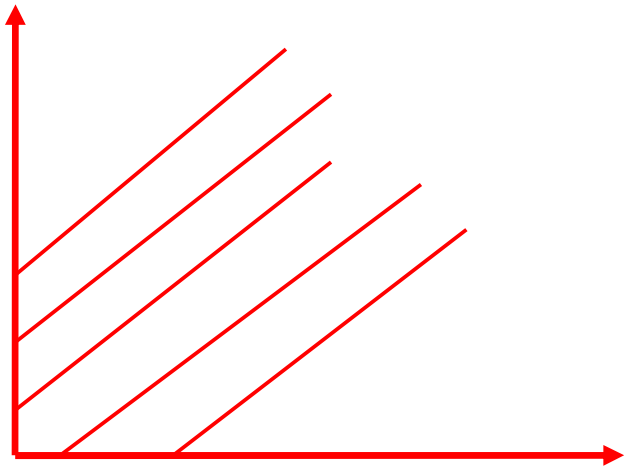
- **SizeZone*Farm Area** (weights constrained to reproduce population and farm area totals in each of six size by zone poststrata)
- **Intercept / no intercept**

Two Random Effects Specifications

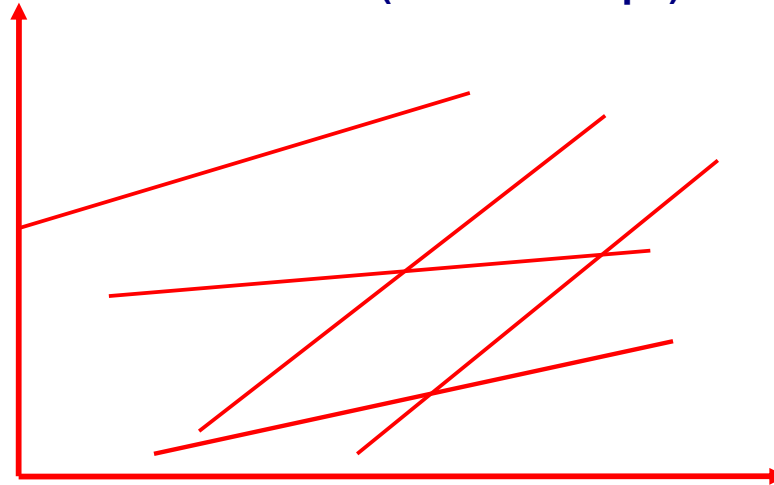
- **Random Intercepts specification** (Z_i equal to a vector on one's)
- **Random Slopes specification** (Z_i equal to the design matrix for a linear regression on Farm Area)

Model Specifications

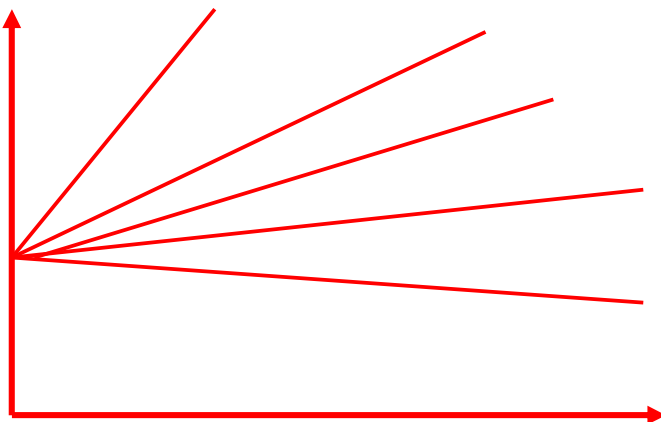
Model -I (Random Intercept)



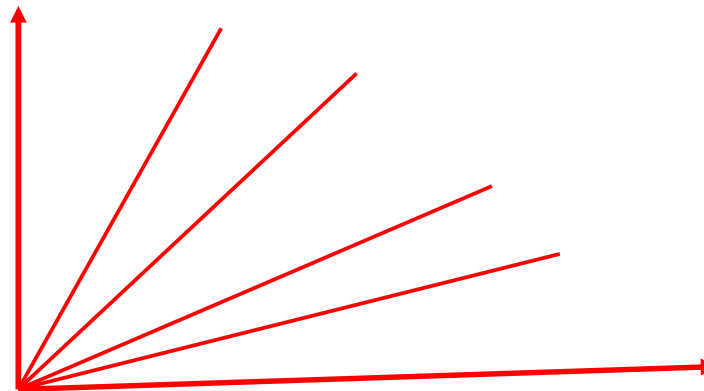
Model -II (Random slope)



Model -III (Random slope and Fixed Intercept)



Model -IV (Random slope and Zero Intercept)



Empirical Results

Average (ARB) and median (MRB) values of relative bias, average (ARRMSE) and median (MRRMSE) values of relative RMSE and average (ACR) coverage rates

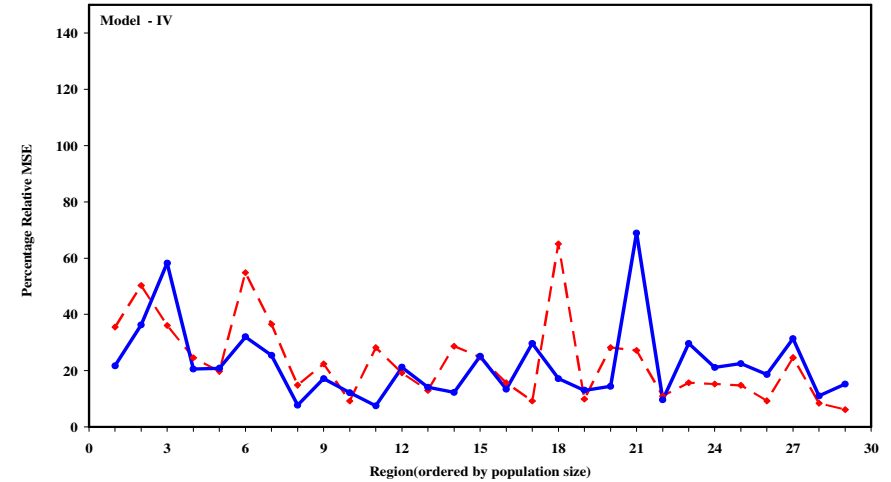
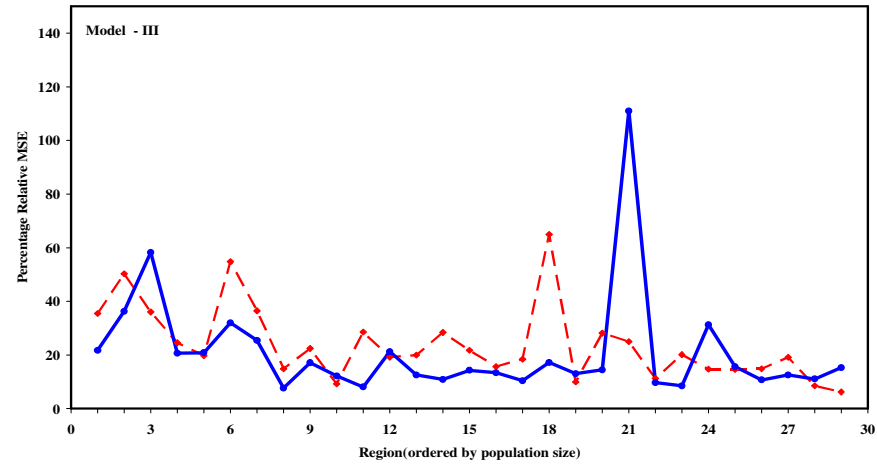
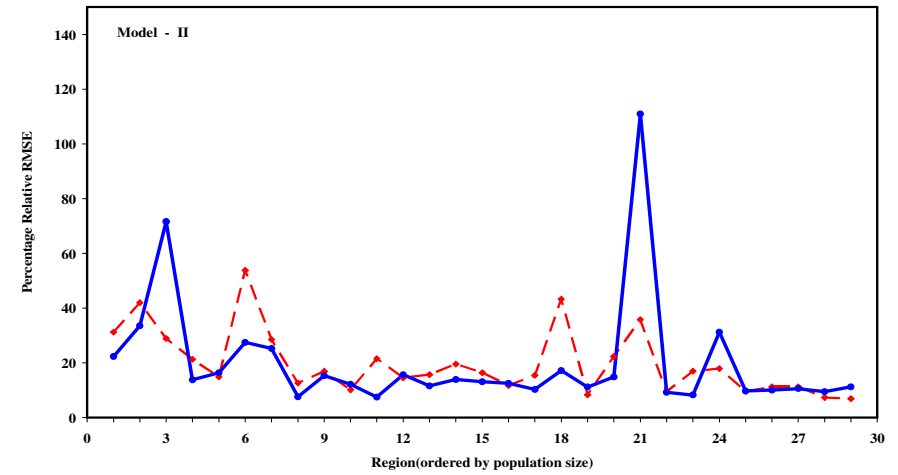
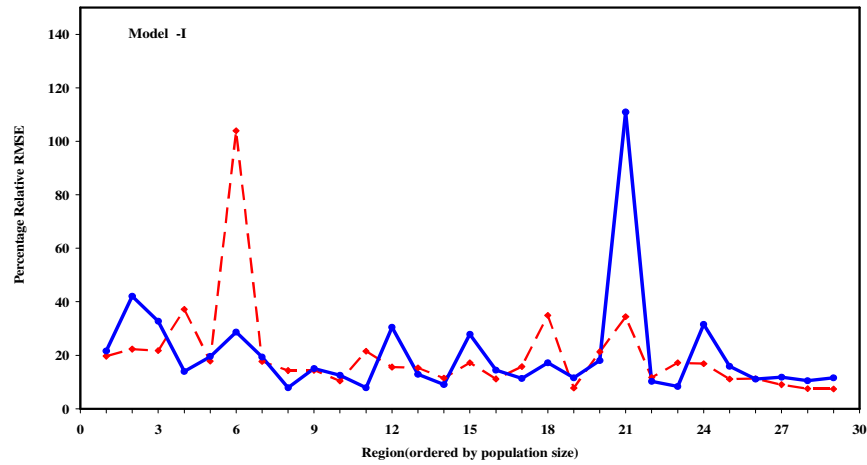
Model	Method	ARB	MRB	ARRMSE	MRRMSE	ACR
I	EBLUP	4.24	1.55	19.92	15.74	0.90
	MBD	-2.49	-0.82	20.56	14.45	0.92
II	EBLUP	2.98	0.61	19.87	16.40	0.85
	MBD	-2.13	-0.47	20.15	13.16	0.93
III	EBLUP	4.52	1.95	23.89	19.94	0.69
	MBD	-3.84	0.13	21.14	14.44	0.94
IV	EBLUP	1.17	-2.63	23.38	19.73	0.65
	MBD	2.20	2.06	22.35	20.61	0.97

- Average and median over 29 small areas

Region-specific Relative Root Mean Squared Errors

—◆— EBLUP

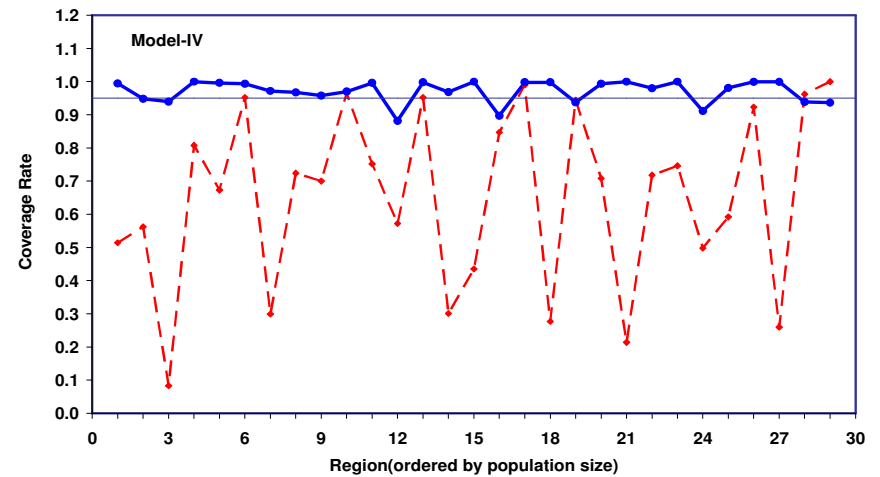
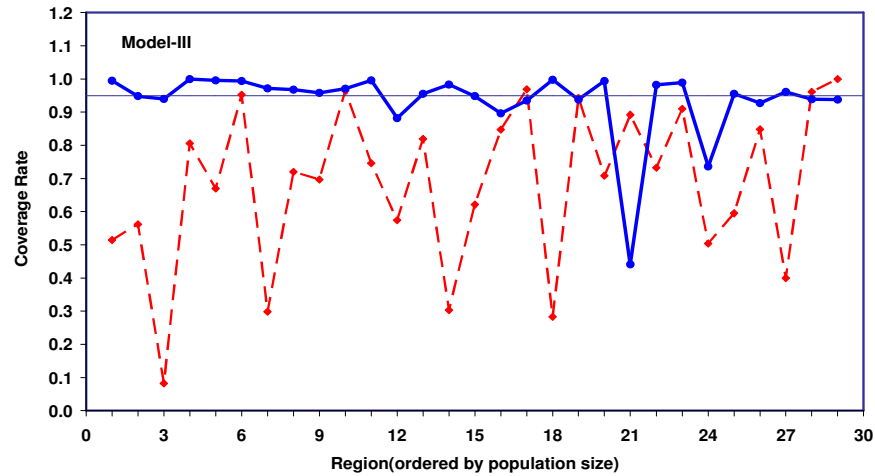
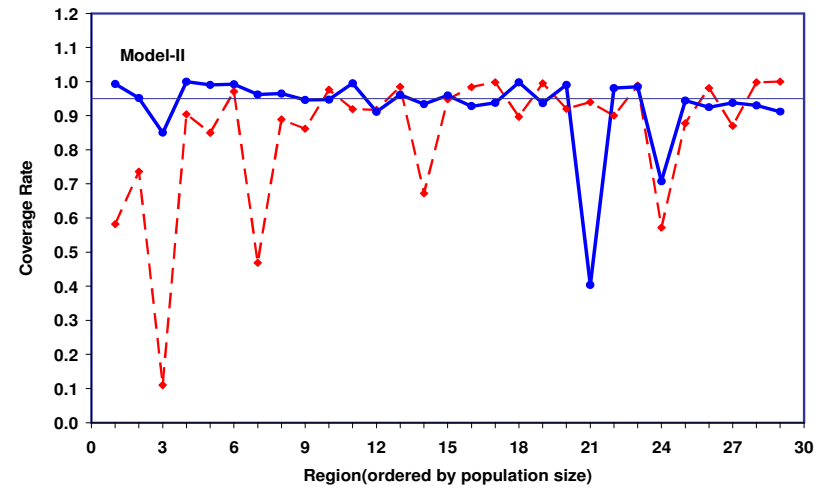
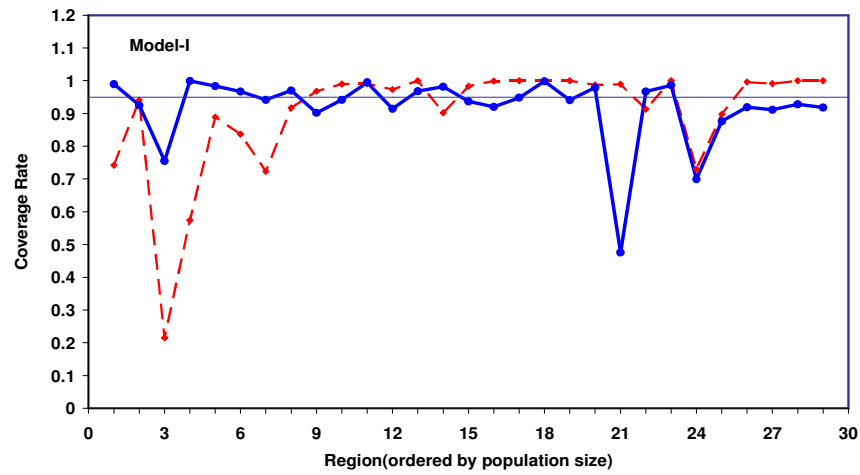
—●— MBD



Region-specific Coverage Rates

—◆— EBLUP

—●— MBD



Conclusions

- In case of model misspecification, the MBD approach appears to provide a more robust set of small area estimates
- The MBD mean squared error estimator performs well and represents a real alternative to the usual EBLUP estimator

References

- Chambers, R. L. and Chandra H (2006). Improved Direct Estimators for Small Areas. In preparation.
- Prasad, N.G.N. and Rao, J.N.K. (1990). The Estimation of the Mean Squared Error of Small Area Estimators. *Journal of the American Statistical Association*, 85, 163-171.
- Rao, J.N.K (2003). *Small Area Estimation*, New York, Wiley.
- Royall, R.M. (1976). The Linear Least Squares Prediction Approach to Two-Stage Sampling. *Journal of the American Statistical Association* 71, 657-664.
- Royall, R.M. and Cumberland, W.G. (1978). Variance Estimation in Finite Population sampling. *Journal of the American Statistical Association* 73, 351-358.