

Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics

Exploration of Approaches to Small Area Estimation with Measurement Errors and their Application to Indonesian Household Surveys

by Ika Yuni Wulansari, Stephen Woodcock, and James Brown

Release date: September 8, 2025



Exploration of Approaches to Small Area Estimation with Measurement Errors and their Application to Indonesian Household Surveys

Ika Yuni Wulansari, Stephen Woodcock, and James Brown¹

Abstract

The United Nations Sustainable Development Goals require detailed, disaggregated data, typically obtained through household surveys. However, surveys alone cannot meet these granular statistics. To address this, National Statistical Institutes adopt small area methods, but these face challenges as auxiliary variables, often derived from surveys, introduce measurement errors into the models. The aim is the application of measurement errors correction in classic Fay-Herriot area-level model. The results demonstrate the robustness of the standard approach and ignoring measurement error but show there are specific scenarios where correction for measurement errors is beneficial. The approach is applied to a case-study utilising Indonesian household survey data.

Key Words: Small area estimation; Granular statistics; Measurement error; EBLUP; Indonesian household data.

1. Introduction and Literature Review

1.1 Small Area Estimation

The development of model-based Small Area Estimation (SAE) in recent years has been very significant. The most popular model used in SAE is the random effect model, which includes a random area effect to explain variations between areas beyond those described by auxiliary variables (Fay III and Herriot, 1979; Rao and Molina, 2015). Based on the availability of auxiliary variables, there are two basic SAE models; area-level and unit-level (Rao and Molina, 2015). The area-level model is applied when auxiliary data is available at an aggregated level, such as districts or cities. The classic area-level random effect model in SAE is the Empirical Best Linear Unbiased Predictor (EBLUP) Fay-Herriot (FH) model, a parametric approach. The alternative unit-level model is used when auxiliary data is accessible for individuals or households. In a standard SAE model, auxiliary variables must be taken from a census or other sources that do not contain measurement errors.

1.2 Small Area Estimation with Measurement Error in Covariates

The success of SAE modelling highly depends on the availability of auxiliary data strongly correlated with the variables to be estimated (Rao and Molina, 2015). However, in many situations, auxiliary information or covariates are unavailable in data sources with full population coverage. It is very likely that other surveys can provide estimates of useful covariates, but with measurement errors. Ybarra and Lohr (2008) extended the FH model to account for sampling variability in the covariates when producing small area estimates. Given estimates of the measurement error variances for the covariates for each small area, the increased uncertainty can be accounted for while still gaining precision from the use of auxiliary information.

¹ Author #1 Ika Yuni Wulansari, University of Technology Sydney, 15 Broadway, Ultimo NSW, Australia, 2007 (ikayuni.wulansari@uts.edu.au), Politeknik Statistika STIS, Indonesia, 13330 (ikayuni@stis.ac.id); Author #2 Stephen Woodcock, University of Technology Sydney, 15 Broadway, Ultimo NSW, Australia, 2007 (stephen.woodcock@uts.edu.au); Author #3 James Brown, University of Technology Sydney, 15 Broadway, Ultimo NSW, Australia, 2007 (james.brown@uts.edu.au)

2. Methodology

2.1 Empirical Best Linear Unbiased Predictor

In this study, we focus on the area-level approach. The model links auxiliary variables $\mathbf{X}_i^T = (X_{1i}, X_{2i}, \dots, X_{pi})$ with direct estimators in each area. The true parameter is θ_i , with a linear model as follows:

$$\theta_i = \mathbf{X}_i^T \boldsymbol{\beta} + u_i, \quad i = 1, 2, \dots, m \quad (1)$$

where $\boldsymbol{\beta} (\beta_0, \dots, \beta_p)^T$ is the $(p+1) \times 1$ vector of regression coefficients; u_i is the small area random effect, assumed $u_i \sim iid N(0, \sigma_u^2)$, and m is number of small areas. The true parameter θ_i is not observed, but from a random sample we have direct estimate $\hat{\theta}_i$ and the sampling model can be written as follows:

$$\hat{\theta}_i = \theta_i + e_i, \quad i = 1, 2, \dots, m \quad (2)$$

where e_i is the sampling error, assumed $e_i \sim iid N(0, \psi_i)$. The combination of (1) and (2) produces the linear mixed model that can be written as follows:

$$\hat{\theta}_i = \mathbf{X}_i^T \boldsymbol{\beta} + u_i + e_i, \quad i = 1, 2, \dots, m \quad (3)$$

The linear mixed model in equation (3) is called the Fay-Herriot (FH) model (Fay and Herriot, 1979), and we can construct the standard FH estimator

$$\tilde{\theta}_i^{BLUP} = \mathbf{X}_i^T \tilde{\boldsymbol{\beta}} + \gamma_i (\hat{\theta}_i - \mathbf{X}_i^T \tilde{\boldsymbol{\beta}}) = \gamma_i \hat{\theta}_i + (1 - \gamma_i) \mathbf{X}_i^T \tilde{\boldsymbol{\beta}}$$

where $\gamma_i = \sigma_u^2 / (\sigma_u^2 + \psi_i)$. This depends on the unknown value of σ_u^2 , but following the standard approach to estimation of the unknown parameters $\boldsymbol{\beta}$ and σ_u^2 (Rao and Molina, 2015) we obtain the Empirical Best Linear Unbiased Predictor (EBLUP) $\hat{\theta}_i^{FH} = \hat{\gamma}_i \hat{\theta}_i + (1 - \hat{\gamma}_i) \mathbf{X}_i^T \hat{\boldsymbol{\beta}}$, where $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \psi_i}$. The unbiased estimator of $MSE(\hat{\theta}_i^{FH})$ follows based on Prasad and Rao (1990).

2.2 Empirical Best Linear Unbiased Predictor with measurement error in covariates

The EBLUP estimator used in SAE analysis assumes the auxiliary variables do not contain an error component. Ybarra and Lohr (2008) extended the EBLUP under the FH model to account for measurement errors in the auxiliary variable, making it possible to use survey data independent to the survey of interest as auxiliary variables, which we refer to as SAE with measurement error (ME). Ybarra and Lohr (2008) used the survey data, $\tilde{\mathbf{X}}_i$, as an estimator of population data, \mathbf{X}_i , which if just inserted in (3) underestimates the MSE. They extended (3) for \mathbf{X}_i with measurement error, to

$$\hat{\theta}_i = \tilde{\mathbf{X}}_i^T \boldsymbol{\beta} + r_i (\tilde{\mathbf{X}}_i, \mathbf{X}_i) + e_i \quad (4)$$

where $r_i(\tilde{\mathbf{X}}_i, \mathbf{X}_i) = u_i + (\mathbf{X}_i - \tilde{\mathbf{X}}_i)^T \boldsymbol{\beta}$. When regression and covariance parameters are known, $C_i = MSE(\tilde{\mathbf{X}}_i)$ is given by $E[(\tilde{\mathbf{X}}_i - \mathbf{X}_i)(\tilde{\mathbf{X}}_i - \mathbf{X}_i)^T]$ and $MSE(r_i) = \sigma_u^2 + \boldsymbol{\beta}^T C_i \boldsymbol{\beta}$. Hence, the BLUP FH equation with measurement error is:

$$\tilde{\theta}_i^{ME} = \gamma_i \hat{\theta}_i + (1 - \gamma_i) \tilde{\mathbf{X}}_i^T \boldsymbol{\beta} \quad (5)$$

with $\gamma_i = \frac{MSE(r_i)}{MSE(r_i) + \psi_i}$ and $MSE(\tilde{\theta}_i^{ME}) = \gamma_i \psi_i$. As above, $\boldsymbol{\beta}$ and σ_u^2 need to be estimated from the data with the assumption that both C_i and ψ_i are known. Ybarra and Lohr (2008) provided the full details of the estimation and once $\boldsymbol{\beta}$ and σ_u^2 are replaced by their consistent estimator $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_u^2$, the EBLUP FH ME estimator is: $\hat{\theta}_i^{ME} = \hat{\gamma}_i \hat{\theta}_i + (1 - \hat{\gamma}_i) \tilde{\mathbf{X}}_i^T \hat{\boldsymbol{\beta}}$, where $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2 + \hat{\boldsymbol{\beta}}^T C_i \hat{\boldsymbol{\beta}}}{\hat{\sigma}_u^2 + \hat{\boldsymbol{\beta}}^T C_i \hat{\boldsymbol{\beta}} + \psi_i}$. Ybarra and Lohr (2008) used the Jackknife method proposed by Jiang et al. (2002) to estimate the $MSE(\hat{\theta}_i^{ME}) = M_{1i} + M_{2i}$, where $M_{1i} = \gamma_i \psi_i$ is the MSE of (5) and $M_{2i} = E(\hat{\theta}_i^{ME} - \tilde{\theta}_i^{ME})^2$ is the additional term due to estimation of the unknown parameters.

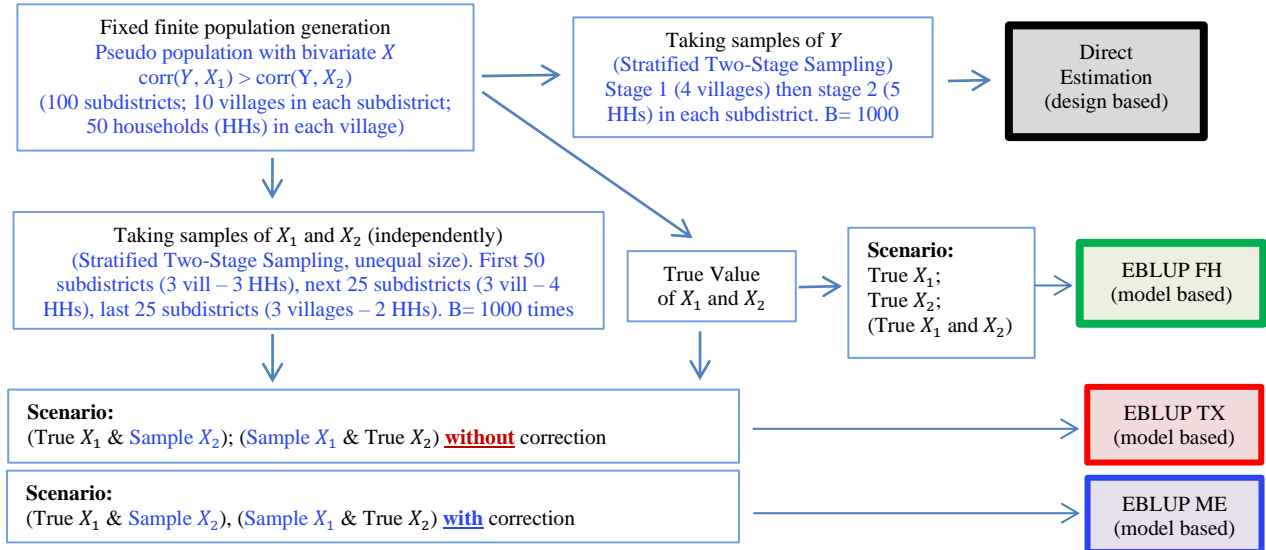
3. Simulation Study

3.1 Simulation procedure

Our simulation framework to evaluate the performance of our estimation methods under different sampling and error correction scenarios is described in Figure 3.1-1. We generated of a fixed finite pseudo-population, structured with 100 subdistricts, 10 villages per subdistrict, and 50 households in each village. This population included bivariate covariates (X_1 and X_2), with Y (the response variable) having a stronger correlation with X_1 than X_2 . Samples of Y are

drawn using stratified two-stage sampling: four villages are selected from each subdistrict at the first stage, followed by five households from each village at the second stage, repeated 1,000 times. Samples of X_1 and X_2 are taken independently using stratified two-stage sampling with unequal sizes across subdistricts, also repeated 1,000 times.

Figure 3.1-1
Simulation Framework



The framework explores scenarios where the true values of X_1 and X_2 are known and contrasts them with cases involving sampled values, both with and without error correction. These scenarios feed into different estimation methods: direct estimation (design-based) relies on Y samples only, while model-based estimators draw on different sources of auxiliary information: (i) EBLUP FH uses true covariate values as auxiliary information, (ii) EBLUP TX evaluates the effect of using sampled X_1 and true X_2 , and vice versa, without corrections, and (iii) EBLUP ME incorporates corrections for measurement errors in the auxiliary information.

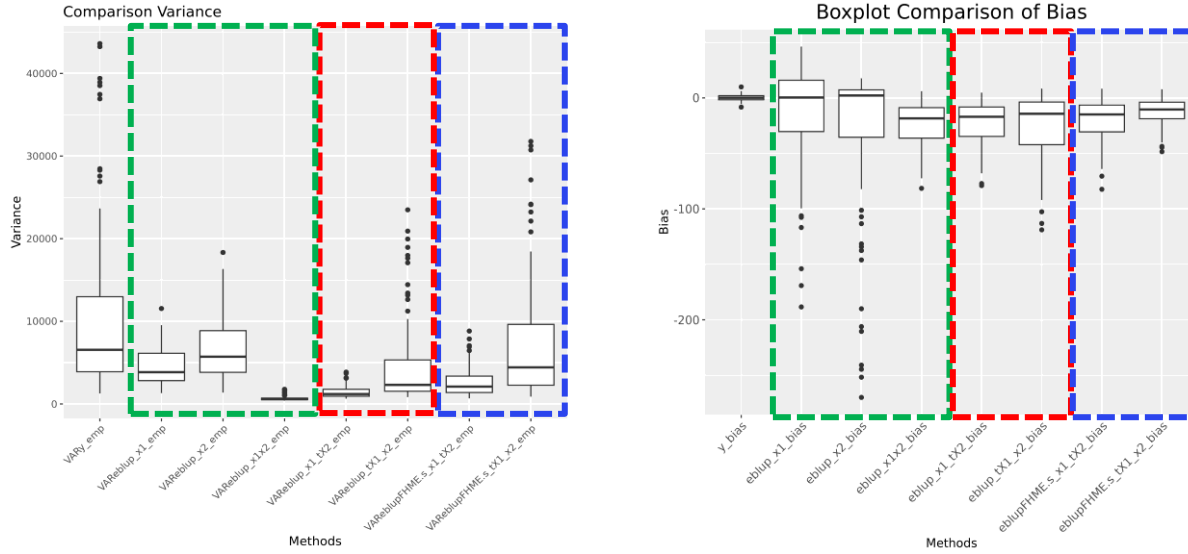
In their evaluation, Ybarra and Lohr (2008) generated their sample data using a model-based approach, while our approach uses a design-based generation to mimic sampling designs found in real data. Ybarra and Lohr (2008) applied $k\%$ of randomly chosen C_i (MSE of the X_i), to equal 3, with the remaining C_i set to 0 (as if there are no errors) for $k \in \{0,20,50,80,100\}$, to simulate varying levels of measurement error across small areas. However, as our approach estimates the covariates from a sample, the sampling errors inherently induce measurement error across all areas, aligning closer to real-world scenarios. Stratification with varying allocations ensures the true measurement error variances C_i still vary across small areas. Furthermore, Ybarra and Lohr (2008) focused on univariate X , while our work incorporates bivariate covariates (X_1 and X_2) to study the impact of different covariate structures on error and variance reduction.

Ybarra & Lohr (2008) simulated univariate cases under four conditions: without error, without a covariate (intercept-only model), with error but no correction, and with error and correction. In contrast, we simulate scenarios under conditions: univariate X without error, bivariate covariates without error (true model), bivariate covariates with error but no correction, and bivariate covariates with error and correction. By including bivariate covariates and focusing on design-based pseudo-population generation, our approach provides a comprehensive and realistic framework for evaluating the impact of covariate structures and sampling designs on small area estimation.

3.2 Simulation result

Figure 3.2-1 presents a comparison of variance and bias between different estimation approaches, demonstrating the trade-offs between variability and accuracy.

Figure 3.2-1
Comparison of Variance (a) and Bias (b)



a. Variance comparison

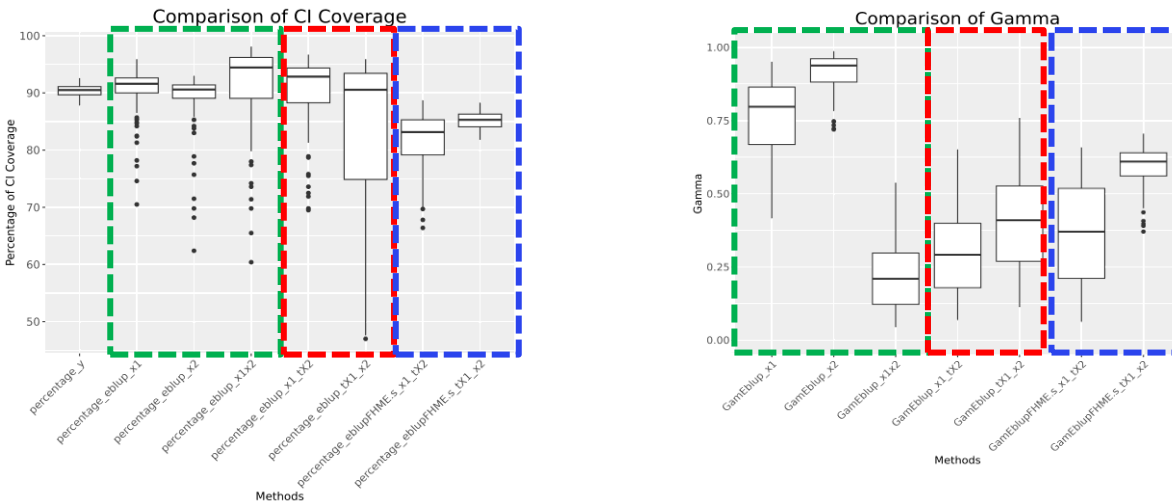
b. Bias comparison

Note:

- Univariate True X_1 – Univariate True X_2 – Bivariate (True X_1 , True X_2), TRUE MODEL
- Bivariate (True X_1 , Sample X_2) – Bivariate (Sample X_1 , True X_2), WITHOUT ME correction
- Bivariate (True X_1 , Sample X_2) – Bivariate (Sample X_1 , True X_2), WITH ME correction

As we would expect, Figure 3.2-1, the EBLUP with bivariate X without error (true model) performs the best in terms of variance reduction across sub-districts, when compared to the direct estimates. EBLUP bivariate X with error (without ME correction) performs better than either univariate X_1 or univariate X_2 . EBLUP bivariate X with error (with ME correction) also performs better than either univariate X_1 or univariate X_2 , but it does not perform better than just ignoring the measurement error. However, the bias plots in Figure 3.2-1 demonstrate correcting the ME offers some protection reducing the distribution of bias across areas.

Figure 3.2-2
Comparison of Confidence Interval (CI) coverage (a) and Gamma (b)



a. Confidence Interval (CI) Coverage comparison

b. Gamma comparison

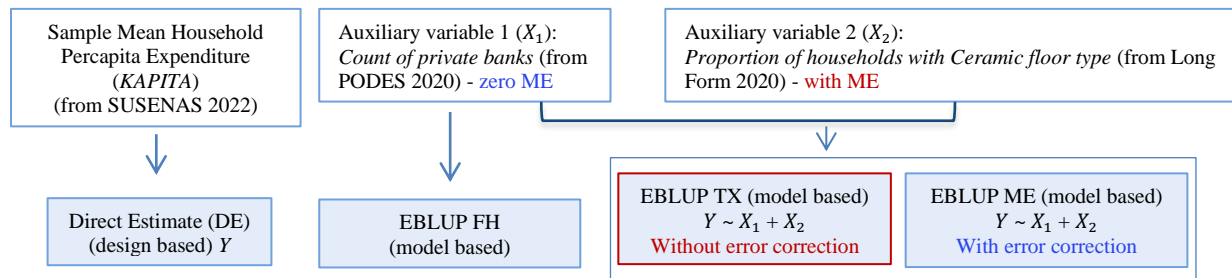
Looking at Figure 3.2-2, as the shrinkage parameter γ of bivariate X gets close to zero, it brings the EBLUP towards the model. However, for ME, shrinkage parameter γ moves closer to one, reflecting the increased variance from measurement error, and this drives the improvement in bias as the estimates take more from the direct estimate. In terms of 95% Confidence Interval (CI) coverage, the EBLUP TX (without ME correction) gives concerning result, reflecting the standard MSE estimator not correcting for the increased uncertainty due to measurement error. The EBLUP ME gives better estimated CI coverage than that of the EBLUP TX.

4. Empirical Study

4.1 Empirical procedure

We applied the EBLUP approach with measurement error correction to data from the Indonesian Household (HH) survey. Figure 4.1-1 illustrated a workflow for empirical study. We estimated the response variable, Y , sample mean of household per capita expenditure (KAPITA), as a design based direct estimate. In the model-based approaches, we estimated the EBLUP FH using X_1 only, which assumed zero measurement error (ME). We also estimated the EBLUP TX model using true X_1 and sampled X_2 , but did not correct for measurement error, and the EBLUP ME model using true X_1 and sampled X_2 , which adjusted for measurement error.

Figure 4.1-1
Empirical framework



The direct estimation process utilized the 2022 National Socio-Economic Survey (SUSENAS), which encompassed 514 districts and 34,468 primary sampling units (PSUs). From this dataset, a subsample of 3,447 PSUs was extracted to replicate subdistrict-level variance. The target variable was the mean of household (HH) per capita expenditure, known as KAPITA, expressed in thousand Rupiahs.

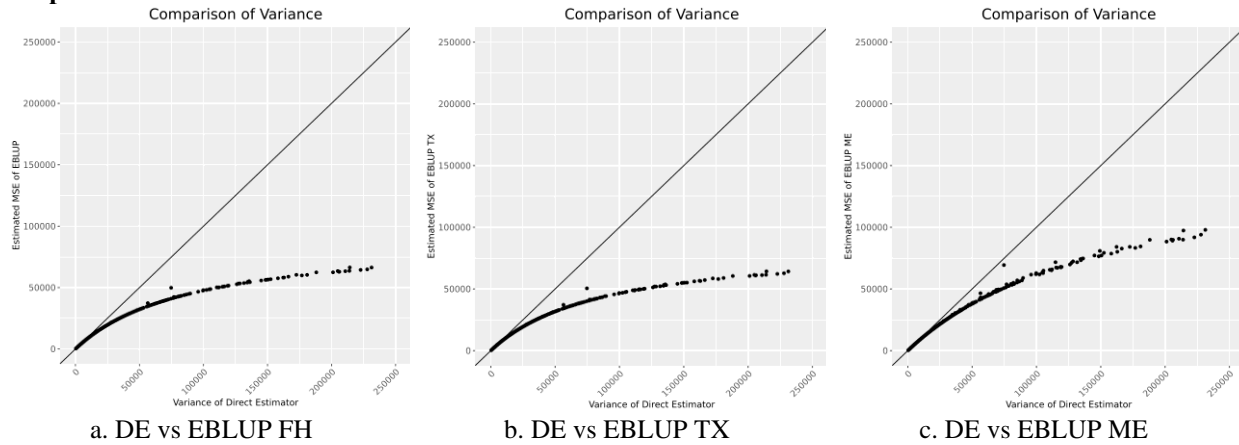
Auxiliary information was sourced from two datasets. The first was the 2020 Village Potential Data (PODES), covering 514 districts and based on administrative records, which assumed zero measurement error. The relevant variable was the count of private banks, which had a correlation with KAPITA. The second source was the 2020 Long Form Survey, which had a wider coverage than SUSENAS, including 514 districts and 268,125 PSUs, which provided auxiliary information with measurement error. From this survey, the proportion of households with ceramic floor types, which had a correlation with KAPITA, was the measurement error (ME) covariate in the model. Since this procedure relied on subsampling, we ensured that this did not change the substance of the original sample. We plotted the subsample against the original sample and verified that the essential characteristics and variance structure of the original sample were preserved.

4.2 Empirical results

Figure 4.2-1 illustrates the comparison between the variances of direct estimators and those of model-based estimators across different scenarios, highlighting the effectiveness of model-based approaches in reducing variance or mean squared error (MSE). The results demonstrate that incorporating the additional variable X_2 with measurement error (ME) correction, referred to as EBLUP ME, yields a slightly less reduction of the MSE from the Direct Estimator

compared to the reduction we got from ignoring the error. This result is consistent with our findings in the simulation, underscoring the trade-off between accounting for errors in auxiliary variables and the resulting efficiency gains.

Figure 4.2-1
Comparison of Variance



5. Conclusion

The EBLUP with bivariate X and no error performs the best in terms of variance reduction. EBLUP bivariate X with error (without ME correction/ EBLUP TX) performs better than univariate X_1 and univariate X_2 in terms of variance and MSE reduction. EBLUP bivariate X with error (with ME correction/ EBLUP ME) also performs better than univariate X_1 and univariate X_2 . Unfortunately, it does not (usually) perform better than just ignoring the error, in terms of variance and MSE reduction. EBLUP ME gives higher gamma, so takes more from the direct estimate (to protect against bias), but less variance reduction. The EBLUP ME also gives better CI coverage than that of the EBLUP without doing ME correction. Using covariates with measurement error correction can still deliver appreciable reduction in bias. Ultimately, SAE is about prediction. There is less focus on the actual β , so traditional measurement error bias appears less important to final estimates.

Acknowledgement

We acknowledge University of Technology Sydney for full financial support to present at the Statistics Canada's International Methodology Symposium 2024. We also acknowledge Indonesia Endowment Fund for Education (LPDP) from the Ministry of Finance Republic Indonesia for the scholarship supporting this research.

References

- Fay III, R. E., and Herriot, R. A. (1979), "Estimates of income for small places: an application of James-Stein procedures to census data", *Journal of the American Statistical Association*, 74(366a), pp. 269-277.
- Jiang, J., Lahiri, P., and Wan, S.-M. (2002), "A Unified Jackknife Theory for Empirical Best Prediction with M-Estimation", *Annals of Statistics*, 30, pp. 1782-1810.
- 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, pp. 163-171.

Rao, J., and Molina, I. (2015), *Small area estimation* (Second edition. ed.). Hoboken, New Jersey: Wiley.
Ybarra, L. M., and Lohr, S. L. (2008), "Small area estimation when auxiliary information is measured with error",
Biometrika, 95(4), pp. 919-931.