Measurement Error Subarea Model: An Application of Farm Labor Parameters

Lu Chen[*,#], Balgobin Nandram[+,#], Linda J. Young[#]

[*] National Institute of Statistical Sciences (NISS)

[#] United States Department of Agriculture
National Agricultural Statistics Service (NASS)

[+] Worcester Polytechnic Institute

FCSM Session Session G-5: Evidence-based Bayesian Methods for More Precise Estimates and Useful Inference
October 23, 2024

Disclaimer and Acknowledgment

The findings and conclusions in this presentation are those of the authors and should not be construed to represent any official USDA or US Government determination or policy.

- This research was supported by the U.S. Department of Agriculture, National Agricultural Statistics Service.

Outline

Motivation

Models

Case Study

Concluding Remarks

Motivation

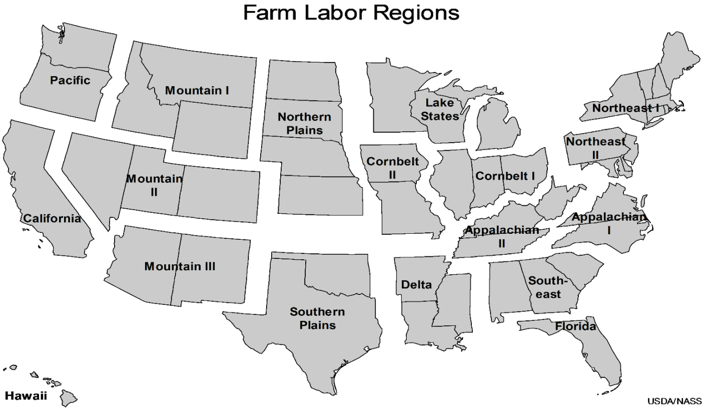
- Hierarchical Bayesian small area models are implemented in many NASS projects including Crops County Estimates, **Farm Labor**, and Cash Rents projects: NASEM (2018, 2023), Young and Chen (2022), Chen, et al. (2022a, 2022b, 2023).
- NASS contracted with NORC to conduct review and research improvements to NASS sampling methods, including for surveys resulting in small area estimation.
- One mid-term (2-4 years) recommendation of NORC's is to consider eliminating, reducing, or accounting for measurement error (ME) in the covariates in the current small area estimation modeling strategies.

Background

- Current models use the previous corresponding year's or quarter's official estimates.
- These covariates are subject to variability that would presumably differ among areas.
- Ignoring measurement error in small area models tends to be particularly problematic when the corresponding variances of the covariates measured with error differ among areas.
- The potential pitfalls include suboptimal prediction and incorrect estimation of uncertainty measures.
- Fuller (2009), Ybarra and Lohr (2007), Arima et al (2017), Bell et al. (2019).

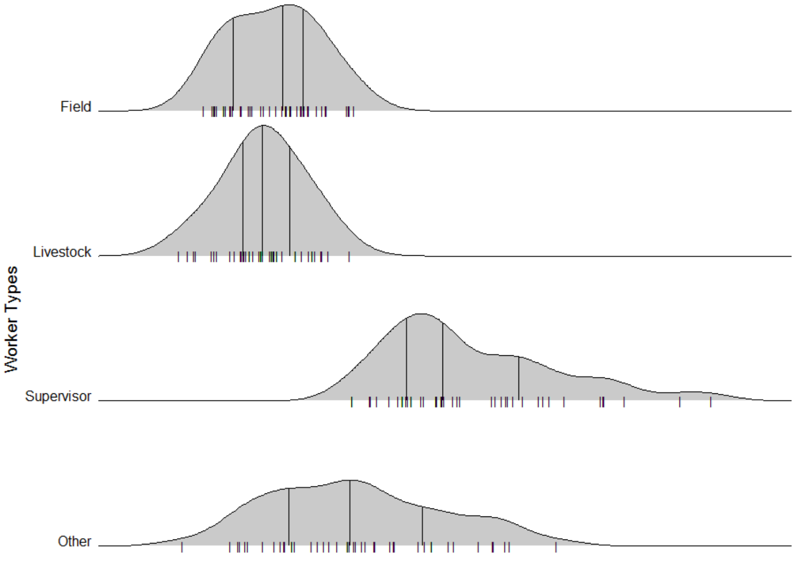
Data: Quantities of Interest

- Regional-level and US-level estimates:



- **NASS Worker Types**; the Standard Occupational Classification (SOC)

Direct Estimates — Wage Rates by Types



Notation

- $i = 1, \dots, m$ index for areas (i.e., regions)
- $j = 1, \dots, n_i$ index for subareas (i.e., states) within area i
- $\hat{\theta}_{ij},\ \hat{\sigma}^2_{ij}$ Farm Labor direct estimates by worker types
- x_{ij} known auxiliary information: the previous year, same quarter, official estimates; number of positive responses; and worker types

Subarea Model for Wage Rates (Original)

The subarea model for wage rates:

$$
\begin{gathered}
\hat{\theta}_{ij}|\theta_{ij} \overset{ind}{\sim} N(\theta_{ij},\ \hat{\sigma}^2_{ij}),\\
\theta_{ij}|\boldsymbol{\beta}, \nu_i, \sigma^2_\mu \overset{ind}{\sim} N(\mathrm{x}'_{ij}\boldsymbol{\beta} + \nu_i, \sigma^2_\mu), j = 1, \ldots, n_i,\\
\nu_i|\sigma^2_\nu \overset{iid}{\sim} N(0,\ \sigma^2_\nu),\ i = 1, \ldots, m,\\
\boldsymbol{\beta} \sim MN(\hat{\boldsymbol{\beta}},\ 1000 \times \hat{\Sigma}_{\hat{\beta}}),\\
\sigma^2_\mu{\sim}\text{Uniform}(R^+),\ \sigma^2_\nu{\sim}\text{Uniform}(R^+),
\end{gathered}
$$

- Goals:
  - State × type wage rate: $y^{wg}_{ijk} = \theta_{ijk}$
  - For publication: regional-level wage rates

$$
y^{wg,(h)}_k = \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} y^{wk,(h)}_{ijk} y^{hr,(h)}_{ijk} y^{wg,(h)}_{ijk}}{\sum_{i=1}^m \sum_{j=1}^{n_i} y^{wk,(h)}_{ijk} y^{hr,(h)}_{ijk}},
$$

where $h = 1, ..., H$ are the draws and K are the worker types.

Conditional Structural Error Subarea Model

- One of the covariate $x_{ij1}(=\theta_{2ij})$ has measurement error, for example, previous estimates.
- Structural error model has non-identifiability issue for parameters.
- Proposed a two-part model and the two parts are connected via the multiplication rule of probability.
- Two subarea models connected by the non-identifiable parameter in the first part of the model:

  $$\pi(\theta_1, \theta_2 | D_1, D_2) = \pi(\theta_1 | \theta_2, D_1)\pi(\theta_2 | D_2),$$

  where D_1 and D_2 are the data from the two parts of the model.
- Given θ_2, all the parameters become identifiable in the first part of the model.

Conditional Structural Error Subarea Model

- The first part of the model $\pi(\theta_1|\theta_2, D_1)$:

$$\hat{\theta}_{1ij}|\theta_{1ij} \overset{ind}{\sim} N(\theta_{1ij},\ \hat{\sigma}^2_{1ij}),$$
$$\theta_{1ij}|\boldsymbol{\beta}_1, \theta_{2ij}, \nu_{1i}, \sigma^2_{\mu_1} \overset{ind}{\sim} N(\mathrm{x}'_{1ij}\boldsymbol{\beta}_1 + \gamma\theta_{2ij} + \nu_{1i}, \sigma^2_{1\mu}),$$

- The second part of the model $\pi(\theta_2|D_2)$:

$$\hat{\theta}_{2ij}|\theta_{2ij} \overset{ind}{\sim} N(\theta_{2ij},\ \hat{\sigma}^2_{2ij}),$$
$$\theta_{2ij}|\boldsymbol{\beta}_2, \nu_{2i}, \sigma^2_{\mu_2} \overset{ind}{\sim} N(\mathrm{x}'_{2ij}\boldsymbol{\beta}_2 + \nu_{2i}, \sigma^2_{2\mu}),$$

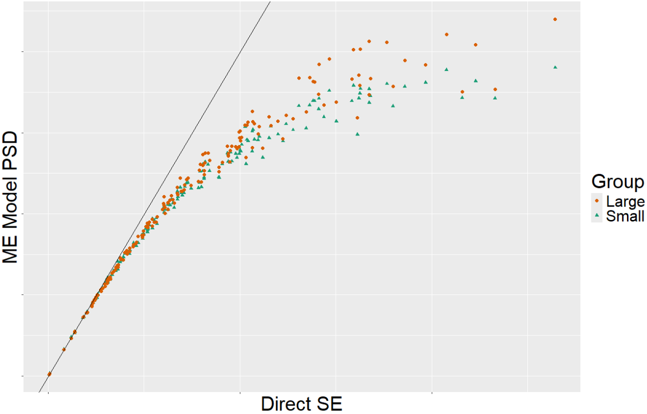
- The priors are similar to the original model.
- Note: $\hat{\theta}_{1ij}$ is the survey estimate and $\hat{\theta}_{2ij}$ is the covariate with measurement errors.

Case Study

- Example:
  - 44 states within 18 regions by worker types
  - Average wage rates
  - Two scenarios of measurement errors are checked:
    - Large variation: previous year's survey variances + noise related to sample sizes
    - Small variation: original model posterior variances based on the previous year survey

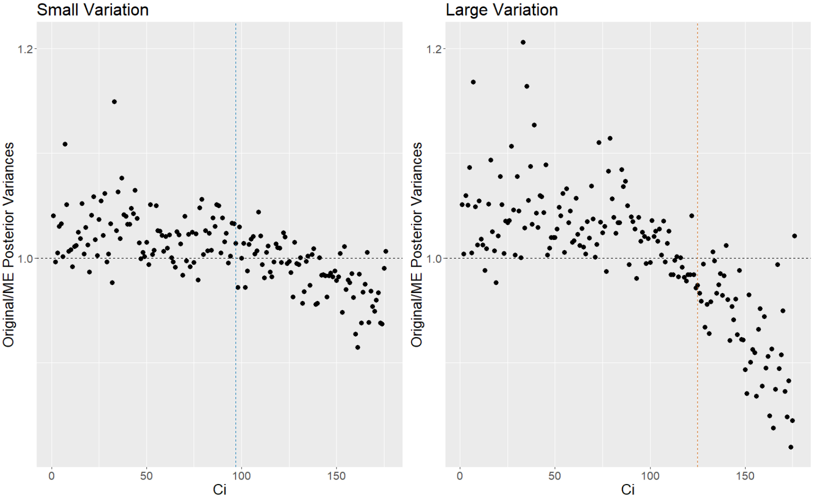
- Computation:
  - 15,000 samples and 5,000 burn-in, 3 chains, each thinned every 10 samples, resulting in a number of 3,000 samples for inference
  - Convergence diagnostics are conducted: Rhat ≤ 1.01 and effective sample sizes are around 3,000

Posterior Standard Deviation Comparisons

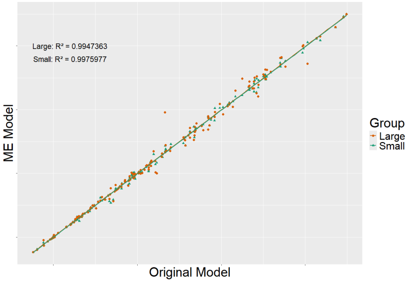


Posterior Variances Ratios v.s. Measurement Errors

Posterior Variances Ratios = Original / ME Posterior Variances



Posterior Mean Comparisons



$$\text{Absolute Relative Differences (\%)} = 100 \times \frac{|ME - Original|}{Original}$$

Cases	Min	25%	Median	Mean	75%	Max
Small	0.002	0.161	0.442	0.682	0.964	4.024
Large	0.005	0.185	0.477	0.962	1.090	14.500

Concluding Remarks

- Investigated the measurement error models from NORC's recommendation
- Proposed the conditional structural error model to avoid the non-identifiablity issue
- The current situation for the previous year's variations are with smaller variations
- However, with large variation, the precision differences are noticeable
- Both posterior means and posterior variances have large differences when the measurement errors are with large variations
- Further research and evaluation are needed

Reference

Arima, S., Datta, G. S., and Liseo, B. (2015).
Bayesian Estimators for Small Area Models when Auxiliary Information is Measured with Error.
Scandinavian Journal of Statistics, 42(2):518–529.

Bell, W. R., Chung, H. C., Datta, G. S., and Franco, C. (2019).
Measurement error in small area estimation: Functional versus structural versus naïve models.
Survey Methodology, 45(1):61–80.

Chen, L., Cruze, N. B., and Young, L. J. (2022a).
Model-based estimates for farm labor quantities.
Stats, 5(3):738–754.

Chen, L. and Nandram, B. (2022).
Combining survey and administrative data to produce official statistics.
In JSM Proceedings, Survey Research Methods Section, pages 1823–1839.

Chen, L., Nandram, B., and Cruze, N. B. (2022b).
Hierarchical bayesian model with inequality constraints for us county estimates.
Journal of Official Statistics, 38(3):709–732.

National Academies of Sciences, Engineering, and Medicine (2018).
Improving Crop Estimates by Integrating Multiple Data Sources.
The National Academies Press.

National Academies of Sciences, Engineering, and Medicine (2023).
Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources.
The National Academies Press.

Ybarra, L. M. R. and Lohr, S. L. (2008).
Small area estimation when auxiliary information is measured with error.
Biometrika, 95(4):919–931.

Young, L. J. and Chen, L. (2022).
Using small area estimation to produce official statistics.
Stats, 5(3):881–897.

Thank You!

lu.chen@usda.gov