Statistical data integration using multilevel models to predict employee compensation

Andreea L. Erciulescu, Jean D. Opsomer, Benjamin J. Schneider, Westat

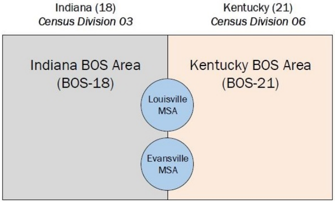
2021 FCSM Research and Policy Conference

November 3, 2021

Contributions

Wage, benefits, and total employee compensation estimates

- Bureau of Labor Statistics (BLS)
- 242,686 domains defined as geography x occupation
  - metropolitan statistical areas (MSAs) and balance of state areas (BOSs); example:



  - 6-digit standard occupational classification codes (SOC6); example:
    - SOC2: 15-0000, Computer and Mathematical Occupations
    - SOC4: 15-2000, Mathematical Science Occupations
    - SOC6: 15-2041, Statisticians

Statistical data integration methodology

- Erciulescu A.L., Opsomer J.D., Schneider, B.J. (2021), "Statistical data integration using multilevel models." Under review.

Data

National Compensation Survey (NCS)

- wage and benefits survey estimates; in \$/hr
  - point estimates: $y_i^{NCS} = (y_{1,i}^{NCS}, y_{2,i}^{NCS})$
  - variance-covariance estimates, adjusted: Σ_i^{NCS}
  - levels: MSA/BOS/census division/nation x SOC6/SOC2/no SOC
  - variations: original scale, log scale, sum
- small sample

Occupational Employment Statistics (OES) program*

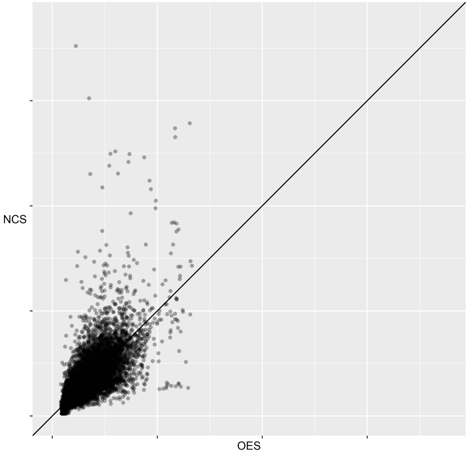
- wage survey estimates; in \$/hr
  - point estimates: $y_{1,i}^{OES}$
  - variance estimates, adjusted: $(\sigma_{1,i}^{OES})^2$
  - levels: MSA/BOS/census division/nation x SOC6/SOC2/no SOC
  - variations: original scale, log scale
- large sample

Prediction space: the set of domains for which there are sample data available in at least one of the two surveys; May 2019 as reference time

Occupational Employment and Wage Statistics Program - as of spring 2021

Need for data integration: distinct wage estimates

Domain-level wage survey estimates, MSA/BOS x SOC6



Two (large) domain-level NCS wage estimates were removed to improve visualization

Need for small area estimation: small sample data

Summary of sample sizes of domains in the prediction space, by level of aggregation; pseudo-effective sample sizes for NCS

Level	NCS			OES		
	Minimum	Median	Maximum	Minimum	Median	Maximum
MSA/BOS x SOC6	0	0	61	0	6	14,826
Census division x SOC6	0	1	191	1	236	68,810
Census division x SOC2	1	49	423	449	11,254	127,475
Nation x SOC6	0	8	796	21	2,272	366,362
Nation x SOC2	7	488	2,208	10,446	112,978	661,453

- median NCS sample size is 1 in NCS-only domains and 1 in all NCS domains
- median OES sample size is 5 in OES-only domains and 6 in all OES domains

Currently, BLS publishes employee compensation statistics at levels of aggregation defined using either geography or occupation (https://www.bls.gov/web/ecec/ececrse.htm).

Need for data integration and small area estimation: incomplete sample data

Number of domains in the prediction space, by level of aggregation

Level	Prediction Space Subset		
	NCS-only	NCS-and-OES	OES-only
MSA/BOS x SOC6	186	19,509	222,991
Census division x SOC6	0	4,358	2,565
Census division x SOC2	0	198	0
Nation x SOC6	0	721	50
Nation x SOC2	0	22	0

- small number of domains with benefits estimates
- large number of domains with two wage estimates
- very large number of domains with wage estimates from only one of the two sources

Hierarchical modeling estimation

Domain-level: MSA/BOS x SOC6-level survey estimates and associated variance estimates

- NCS-only domains (s_{NCS}), NCS-and-OES domains ($s_{NCS-OES}$), OES-only domains (s_{OES})

Bivariate: wage and benefits

- borrow strength from the strong relationship

Hierarchical Bayes: sampling levels, smoothing (latent) level, prior distributions

- borrow strength across surveys, across domains, and from covariates
  - covariates x_i defined in terms of area type (MSA or BOS), census division, and their two-way interactions
- link the NCS and OES wage estimates
- maintain the relationship between wage and benefits

Multi-fold: MSA/BOS x SOC6, SOC6

- borrow strength from the nested structure

Domain-level bivariate hierarchical Bayes multi-fold model

Sampling Level

$$\begin{aligned} y^{NCS}_{i,log} | (\theta_{i,log}, \Sigma^{NCS}_{i,log}) &\sim \mathrm{N}(\theta_{i,log}, \Sigma^{NCS}_{i,log}), i \in s_{NCS} \cup s_{NCS-OES} \\ y^{OES}_{1,i,log} | (\theta_{1,i,log}, \sigma^{OES}_{1,i,log}) &\sim \mathrm{N}\left(\theta_{1,i,log}, (\sigma^{OES}_{1,i,log})^2\right), i \in s_{OES} \cup s_{NCS-OES} \end{aligned}$$

Smoothing Level

$$\begin{aligned} \theta_{i,log} | (\beta, u_I, \Sigma_b) &\sim \mathrm{N}(x_i'\beta + u_I, \Sigma_b), i \in s_{NCS} \cup s_{NCS-OES} \cup s_{OES}, i \in I \\ u_I | \Sigma_u &\sim \mathrm{N}(0, \Sigma_u), i \in s_{NCS} \cup s_{NCS-OES} \cup s_{OES}, i \in I \end{aligned}$$

Prior Distributions

$$\begin{aligned} \beta &\sim \mathrm{N}(0, 10^4), \text{ component-wise} \\ (\Sigma_b, \Sigma_u) &\sim \text{inverse-Wishart}(I_2, 3), \text{ component-wise} \end{aligned}$$

- i indexes MSA/BOS x SOC6 domains
- I indexes SOC6 domains

Model fit, assumptions checks, prediction

Fit

- R JAGS
- Markov chain Monte Carlo (MCMC): 3 chains, 10,000 samples, 3,000 burn-in, thinning every 10th sample: 2,100 samples for inference
- SOC2-specific: 22 models

Assumptions checks

- MCMC diagnostics: $\hat{R}$, MC effective sample size, MC standard error, autocorrelation
- model specification: posterior predictive checks

Prediction

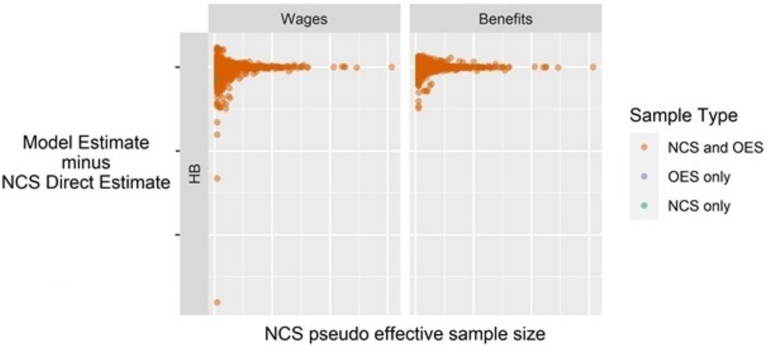
- posterior distribution

  $$[\theta_{i,log} | y_{log}^{NCS}, y_{1,log}^{OES}, \Sigma_{log}^{NCS}, \sigma_{1,log}^{OES}, x, \beta, \Sigma_b, \Sigma_u], i \in s_{NCS} \cup s_{NCS-OES} \cup s_{OES}$$

- transformations: exponential, sum

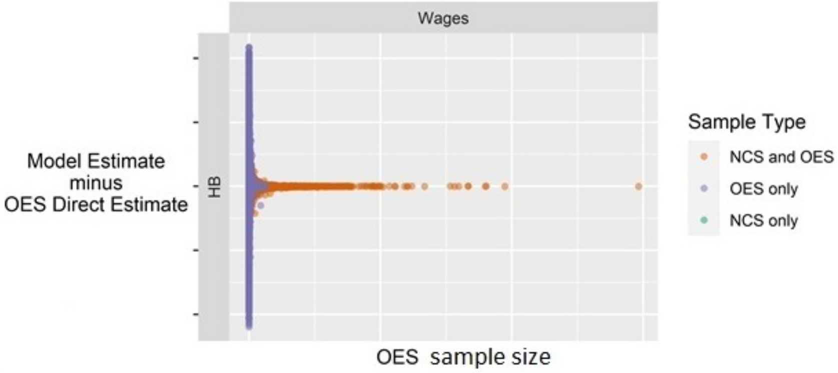
Comparison of NCS and model: point estimates

Domain-level wage and benefits estimates, MSA/BOS x SOC6



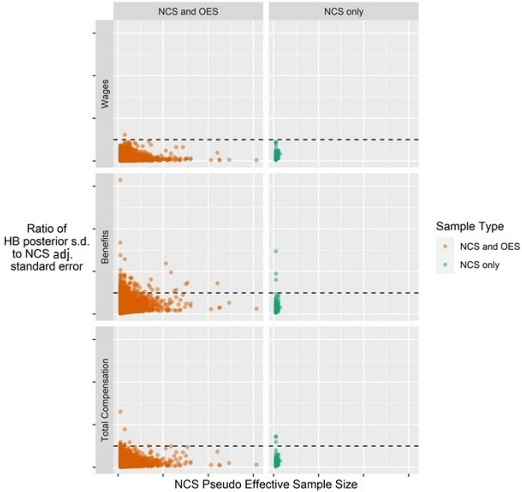
Comparison of OES and model: point estimates

Domain-level wage and benefits estimates, MSA/BOS x SOC6



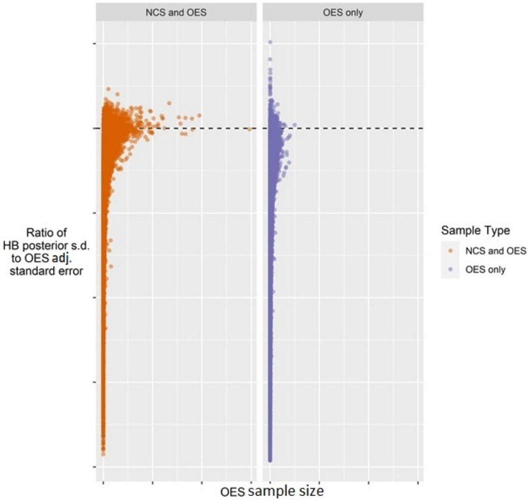
Comparison of NCS and model: standard errors

Domain-level wage and benefits estimates, MSA/BOS x SOC6



Comparison of OES and model: standard errors

Domain-level wage and benefits estimates, MSA/BOS x SOC6



Comparison of NCS, OES, and model: coefficients of variation

Summary of coefficients of variation (%) of compensation estimates for the MSA/BOS x SOC6 domains in the prediction space

Estimation Approach	Wages		Benefits		Total Compensation	
	Median	$\% \geq 30$	Median	$\% \geq 30$	Median	$\% \geq 30$
Survey, NCS; adj. s.e.	49	77	90	92	58	83
Survey, OES; adj. s.e.	17	27	N/A	N/A	N/A	N/A
Model, HB	9	0	28	44	11	1

Recall there are 242,686 domains in the prediction space

Summary

- Methodological developments in statistical data integration, as extensions to small area estimation
- Incomplete survey data on two strongly-related variables
  - one variable collected on two surveys, the other collected only on the smaller survey
  - domains of interest represented by the union of the domains with sample data available for either variable and from either survey
- Complete set of wage, benefits, and total compensation estimates for all domains of interest, with associated uncertainty measures
  - granular levels lower than the levels at which current official statistics are available
- Hierarchical model estimates of improved precision, compared to the survey direct estimates

Selected references

Erciulescu, A.L., and J.D. Opsomer. 2019. "Task Order 5: Developing a Small Domain Estimation Methodology for the Office of Compensation and Working Conditions: Subtask 9: Final Report." Report Prepared for Bureau of Labor Statistics' Office of Compensation and Working Conditions.

Goodman, L. A. 1960. "On the Exact Variance of Products." Journal of the American Statistical Association 55 (292): 708-13.

Guciardo, C. J. 2001. "Estimating Variance in the National Compensation Survey, Using Balanced Repeated Replication." Accessed November 9, 2020. https://www.bls.gov/osmr/research-papers/2001/pdf/st010110.pdf

Hájek, J. 1971. "Comment on a paper by D. Basu." Foundations of statistical inference, 236.

Lettau, M. K., and D. A. Zamora. 2013. "Wage estimates by job characteristic: NCS and OES program data." Monthly Labor Review, U.S. Bureau of Labor Statistics, August. https://doi.org/10.21916/mlr.2013.27.

Myers, M., and D. A. Zamora. 2015. "Revisiting the Dilemma of Review for Modeled Wage Estimates by Job Characteristic." Monthly Labor Review, U.S. Bureau of Labor Statistics, September. https://doi.org/10.21916/mlr.2015.36.

Thank you!

AndreeaErciulescu@westat.com

JAGS two-fold model specification

```
model{
  for(i in 1:mNCS){
        thetahatiNCS[i,1:C] ~ dmnorm(thetai12[i,1:C],  vhatdiriNCS.inv[i,1:C,1:C])
        vhatdiriNCS.inv[i,1:C,1:C] = inverse(vhatdiriNCS[i,1:C,1:C])
  }

  for(i in (mNCS0+1):m){
    thetahatiOES[i] ~ dnorm(thetai12[i,1],  vhatdiriOES.inv[i])
    vhatdiriOES.inv[i] = inverse(vhatdiriOES[i])
  }

  for(i in 1:m){
    thetai12[i,1] =  X1[i,1:P1]%*%beta1[1:P1] + v[i,1] + u[soc6s[i],1]
    thetai12[i,2] = X2[i,1:P2]%*%beta2[1:P2] + v[i,2] + u[soc6s[i],2]
    v[i,1:C] ~  dmnorm(muv[1:C], sigma2v.inv[1:C,1:C])
  }

  for (i in 1:mSOC6s){
    u[i,1:C] ~  dmnorm(muu[1:C], sigma2u.inv[1:C,1:C])
  }

  ## Priors:
  for (p in 1:P2){
      beta2[p] ~ dnorm(0, 1/100)
  }
  for (p in 1:P1){
    beta1[p] ~ dnorm(0, 1/100)
  }

  sigma2v.inv ~ dwish(Kv, 3)
  sigma2v = inverse(sigma2v.inv)

  sigma2u.inv ~ dwish(Ku, 3)
  sigma2u = inverse(sigma2u.inv)
}
```