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

Data

National Compensation Survey (NCS)
- wage and benefits survey estimates; in $/hr
  - point estimates: $y^{NCS}_i = (y^{NCS}_{1,i}, y^{NCS}_{2,i})$
  - variance-covariance estimates, adjusted: $\Sigma^{NCS}_i$
  - 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: $y^{OES}_{1,i}$
  - variance estimates, adjusted: $(\sigma^{OES}_{1,i})^2$
  - 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 $\times$ 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

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

- 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/ecce/ecedcrse.htm).
Need for data integration and small area estimation: incomplete sample data

<table>
<thead>
<tr>
<th>Level</th>
<th>Prediction Space Subset</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NCS-only</td>
<td>NCS-and-OES</td>
<td>OES-only</td>
</tr>
<tr>
<td>MSA/BOS × SOC6</td>
<td>186</td>
<td>19,509</td>
<td>222,991</td>
<td></td>
</tr>
<tr>
<td>Census division × SOC6</td>
<td>0</td>
<td>4,358</td>
<td>2,565</td>
<td></td>
</tr>
<tr>
<td>Census division × SOC2</td>
<td>0</td>
<td>198</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Nation × SOC6</td>
<td>0</td>
<td>721</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Nation × SOC2</td>
<td>0</td>
<td>22</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

- 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 × 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 × SOC6, SOC6
  ▶ borrow strength from the nested structure
Domain-level bivariate hierarchical Bayes multi-fold model

Sampling Level

\[ y_{\text{NCS}}^i \mid (\theta_i, \log, \Sigma_{i,\text{log}}) \sim N(\theta_i, \log, \Sigma_{i,\text{log}}), i \in s_{\text{NCS}} \cup s_{\text{NCS} - \text{OES}} \]

\[ y_{\text{OES}}^i \mid (\theta_1, i, \log, \sigma_{1, i, \log}) \sim N \left( \theta_1, i, \log, (\sigma_{1, i, \log})^2 \right), i \in s_{\text{OES}} \cup s_{\text{NCS} - \text{OES}} \]

Smoothing Level

\[ \theta_{i, \log} \mid (\beta, u_I, \Sigma_b) \sim N(x_i' \beta + u_I, \Sigma_b), i \in s_{\text{NCS}} \cup s_{\text{NCS} - \text{OES}} \cup s_{\text{OES}}, i \in I \]

\[ u_I \mid \Sigma_u \sim N(0, \Sigma_u), i \in s_{\text{NCS}} \cup s_{\text{NCS} - \text{OES}} \cup s_{\text{OES}}, i \in I \]

Prior Distributions

\[ \beta \sim N(0, 10^4), \text{ component-wise} \]

\[ (\Sigma_b, \Sigma_u) \sim \text{inverse-Wishart}(I_2, 3), \text{ component-wise} \]

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

![Graph showing comparison of OES and model estimates](image-url)
Comparison of NCS and model: standard errors

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

![Graph showing comparison between NCS and OES estimates for wages, benefits, and total compensation.](image-url)
Comparison of OES and model: standard errors

Domain-level wage and benefits estimates, MSA/BOS × 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

<table>
<thead>
<tr>
<th>Estimation Approach</th>
<th>Wages</th>
<th>Benefits</th>
<th>Total Compensation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median % ≥ 30</td>
<td>Median % ≥ 30</td>
<td>Median % ≥ 30</td>
</tr>
<tr>
<td>Survey, NCS; adj. s.e.</td>
<td>49 77</td>
<td>90 92</td>
<td>58 83</td>
</tr>
<tr>
<td>Survey, OES; adj. s.e.</td>
<td>17 27</td>
<td>N/A N/A</td>
<td>N/A N/A</td>
</tr>
<tr>
<td>Model, HB</td>
<td>9 0</td>
<td>28 44</td>
<td>11 1</td>
</tr>
</tbody>
</table>

Recall there are 242,686 domains in the prediction space
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


Thank you!

AndreeaErciulescu@westat.com
JAGS two-fold model specification

model{
    for(i in 1:mNCS){
        theta1hatiNCS[i,1:C] ~ dnorm(theta12[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 (mNCS+1):m){
        theta1hatiOES[i] ~ dnorm(theta12[i,1], vhatdiriOES.inv[i])
        vhatdiriOES.inv[i] = inverse(vhatdiriOES[i])
    }

    for(i in 1:m){
        theta12[i,1] = X1[i,1:P1]*%*%beta1[1:P1] + v[i,1] + u[soc6s[i],1]
        theta12[i,2] = X2[i,1:P2]*%*%beta2[1:P2] + v[i,2] + u[soc6s[i],2]
        v[i,1:C] ~ dnorm(muv[1:C], sigma2v.inv[1:C,1:C])
    }

    for (i in 1:mSOC6s){
        u[i,1:C] ~ dnorm(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)