Hierarchical models in the production of official statistics: a discussion of some practical aspects

Andreea L. Erciulescu

Westat

GASP, Washington, DC
September 23, 2019

The views presented in this paper are those of the author(s) and do not represent the views of any Government Agency/Department or Westat

Why hierarchical models in the production of official statistics?

- need to account for various sources of error
  - sampling error, measurement error, linking error
- need for a transparent, reproducible and validated process
  - analytic techniques
- need for measures of uncertainty
  - entire distribution

Hierarchical model - general form

$$\text{Data model:} \quad [y|\theta,\psi]$$

$$\text{Process model:} \quad [\theta|\psi]$$

$$\text{Parameter model:} \quad [\psi]$$

Joint distribution

$$[y,\theta,\psi] = [y|\theta,\psi][\theta|\psi][\psi]$$

Predictive distribution

$$[\theta|y,\psi]$$

Three cases

- known ψ
- unknown, fixed ψ
- unknown, random ψ

Why model-based small area estimation (SAE) in the production of official statistics?

- need to account for various sources of error
- need for a transparent, reproducible and validated process
- need for measures of uncertainty
- integration of data from multiple sources
  - observed survey data within area + auxiliary data within area + information across all areas
- growing demand for granular statistics
  - quantities of interest: i.e. totals, means, ratios
  - 'area' / 'domain': i.e. geography, socio-economic status, occupation
- increasing costs of data collection
  - 'small': amount of survey data available for estimation within a given area (realized sample size as small as zero)

Small domain hierarchical model - general form

Sampling model: $[y_i | \theta_i, x_i, \psi_{\theta,i}, \psi_{y,i}]$

Linking model: $[\theta_i | x_i, \psi_\theta]$

Parameter model: $[\psi_{\theta,i}, \psi_{y,i}]$

Joint distribution

$$[y_i, \theta_i, x_i, \psi_\theta, \psi_{y,i}] = [y_i | \theta_i, x_i, \psi_\theta, \psi_{y,i}][\theta_i | x_i, \psi_\theta][\psi_\theta, \psi_{y,i}]$$

Predictive distribution

$$[\theta_i | y_i, x_i, \psi_\theta, \psi_{y,i}]$$

- *typically* known $x_i, \psi_{y,i}$ (area-level models)
- still three cases
  - known ψ_θ
  - unknown, fixed ψ_θ
  - unknown, random ψ_θ

Examples of model-based SAE in government programs

- Chilean Ministerio de Desarrollo, World Bank: poverty mapping
  - Casas Cordero Valencia et al. (2016); Elbers et al. (2003)
- Census Bureau: income and poverty measures
  - Bell et al. (2016)
- National Agricultural Statistics Service:
  - cash rental rates, Erciulescu et al. (2018)
  - crops production, Erciulescu et al. (2019a)
  - agricultural labor wages, Erciulescu (2018)
- Organisation for Economic Co-operation and Development: adult competency
  - Krenzke et al. (2019)
- Bureau of Labor Statistics: employee compensation components
  - Erciulescu and Opsomer (2019)

A small domain hierarchical Bayes multivariate model for employee compensation components*

$$\text{Sampling Model:} \qquad y_i|\theta_i \sim N(\theta_i, \Sigma_{ei})$$

$$\text{Linking Model:} \qquad \theta_i|(\beta, \Sigma_v) \sim N(x_i\beta, \Sigma_v)$$

$$\text{Independent priors:} \qquad \pi(\beta, \Sigma_v) = \pi(\beta)\pi(\Sigma_v)$$

- domain i, cross-tabulation of census divisions, 6-digit SOC system codes, work levels, binary characteristics
- θ_i, quantities of interest, wage and benefits
- y_i, direct survey estimates, domain-level wage and benefits direct survey estimates
- x_i, known covariates, selected using sample domains definitions
- β, regression coefficients
- Σ_v, linking model variance-covariance matrix
- Σ_{ei}, known survey variance-covariances matrices

* Erciulescu and Opsomer, 2019

A common approach for fit and prediction

Fit

- Markov chain Monte Carlo (MCMC): multiple chains, iterations, burn-in, thining; keep R samples for inference

Prediction

- in-sample domain i
  - R samples $\theta_{i\zeta}, \zeta = 1, \cdots, R$, from $[\theta_i | (y_i, \psi_\theta, \psi_{y,i})]$
- not-in-sample domain $i^{'}$:
  - generate new R samples $\theta_{i^{'}\zeta}, \zeta = 1, \cdots, R$, from $[\theta_i | x_i, \psi_{\theta,\zeta}]$
- small domain posterior means: $R^{-1} \sum_{\zeta=1}^{R} \theta_{i\zeta}$
- small domain posterior variances: $R^{-1} \sum_{\zeta=1}^{R} \left(\theta_{i\zeta} - R^{-1} \sum_{\zeta=1}^{R} \theta_{i\zeta} \right)^2$
- small domain posterior p quantiles: $\boldsymbol{\theta}_{i(p)}, \boldsymbol{\theta}_i = (\theta_{i1}, \cdots, \theta_{iR})$

Practical challenges: number of MCMC samples, large number of domains (in-sample and not-in-sample), variable selection

Prediction for other functions

For example,

- sum of θ_i components (total employee compensation, proportions of adult literacy/numeracy scores in prespecified ranges)
  - inference using R samples $\theta^S_{i\zeta} := \theta_{i\zeta,1} + \theta_{i\zeta,2}, \zeta = 1, \cdots, R$
- ratio of θ_i components (labor wage as ratio of total income to total hours worked, crop yield as the ratio of total production to total harvested acres)
  - inference using R samples $\theta^R_{i\zeta} := \theta_{i\zeta,1}/\theta_{i\zeta,2}, \zeta = 1, \cdots, R$
- aggregates of θ_i for pre-defined domains (2-digits SOC codes, county-agricultural district-state-census division-nation)
  - inference using R samples $\theta^D_{i\zeta} := \sum_{i \in \text{pre-defined domain}} \theta_{i\zeta}, \zeta = 1, \cdots, R$

Examples of model validation

Internal

- mixing and convergence diagnostics: $\hat{R}$, MC effective sample size, autocorrelation
- residuals diagnostics: unconditional and conditional
- posterior predictive checks: indicator, correlation, deviance, residuals
- alternative model specifications: prior distributions

External

- predictions versus direct estimates
- predictions for not-in-sample domains versus predictions for in-sample domains
- cross-validation

Practical challenges: autocorrelation, cross-validation, visualization, storage

Simulation study: Data generation model

$$
\begin{aligned}
y_i|\theta_i &\sim N(\theta_i, \Sigma_{ei}) \\
\theta_i|(\beta, \Sigma_v) &\sim N(x_i\beta, \Sigma_v) \\
\pi(\beta, \Sigma_v) &= \pi(\beta)\pi(\Sigma_v)
\end{aligned}
$$

- $i = 1, \cdots, m$
- $y_i = (y_{i,1}, y_{i,2})$
- x_i, two identical rows $x_{i,row} = (x_{i,1}, x_{i,2}, x_{i,3}, \ldots, , x_{i,(p+1)})$
  - $x_{i,1} = 1$, $x_{i,2} \sim Beta(2, 4)$, $x_{i,k} \sim N(\mu_{xk}, \sigma^2_{xk}), k = 3, \ldots, (p+1)$
  - $\mu_{xk} \sim Unif(1, 50), \sigma_{xk} \sim Unif(1, 10), k = 3, \ldots, (p+1)$
- $\beta = (\beta_1, \beta_2)$, $\beta_j = (\beta_{j,1}, \beta_{j,2}, \beta_{j,3}, \ldots, , \beta_{j,(p+1)}), j = 1, 2$
  - $\beta_{1,1} = 1$, $\beta_{1,2} = 10$, $\beta_{1,k} \sim Unif(1, 5) + N(0, 1), k = 3, \ldots, (p+1)$
  - $\beta_{2,1} = 1$, $\beta_{2,2} = 10$, $\beta_{2,k} \sim Unif(1, 2) + N(0, 1), k = 3, \ldots, (p+1)$
- $\Sigma_v \sim inverse - Wishart(3, I_2)$
- $diag(\Sigma_{ei}) = (\sigma^2_{ei,1}, \sigma^2_{ei,2})$, $\sigma^2_{ei,j} = \exp(\log(x_{i,j}\beta_j) + N(0, 1)), j = 1, 2$
- $cor(y_{i,1}, y_{i,2}) \sim Unif(0.3, 0.9)$

Simulation study: Fitted models

$$
\begin{aligned}
y_i|\theta_i &\sim N(\theta_i, \Sigma_{ei}) \\
\theta_i|(\beta, \Sigma_v) &\sim N(x_i\beta, \Sigma_v) \quad \leftrightarrow \quad \theta_i|(\beta, v_i) = x_i\beta + v_i, v_i|\Sigma_v \sim N(0, \Sigma_v) \\
\pi(\beta, \Sigma_v) &= \pi(\beta)\pi(\Sigma_v)
\end{aligned}
$$

Software

- R STAN, R JAGS

Bayesian specification

- hierarchical Bayes: $\pi(\beta) = N(0, 10^4)$ and $\Sigma_v \sim Inverse - Wishart(3, I)$
- empirical Bayes: $\pi(\beta_1) = N(0, 10^4)$; β_{-1} set equal to the least squares estimates, based on a multiple linear regression model and $\Sigma_v \sim Inverse - Wishart(3, I)$

Practical challenges: prior distributions for the linking model variance-covariance components (Inverse-Gamma/Uniform/Cauchy/F; Inverse-Wishart/LKJ)

Simulation study: Fitted models specifications

Table 1: Model Specifications

Parameters (m, p)	Model	Software	# MC samples / chain (start, burn-in, thin)
$(100, 10)$	S1	STAN	(3000, 1000, 10)
	J1	JAGS	(3000, 1000, 10)
	J1l	JAGS	(30000, 10000, 10)
$(1000, 100)$	S2	STAN	(3000, 1000, 10)
	J2	JAGS	(3000, 1000, 10)
	J2l	JAGS	(30000, 10000, 10)
$(10000, 100)$	S3	STAN	(3000, 1000, 10)
	J3	JAGS	(3000, 1000, 10)
	J3l	JAGS	(30000, 10000, 10)

Simulation study: JAGS HB model specification

```
model{
      for(i in 1:m){
        y[i,1:C] ~ dmnorm(theta[i,1:C],  Sigmae.inv[i,1:C,1:C])
        theta[i,1:C] = thetai0[i,1:C] + v[i,1:C]
        v[i,1:C] ~  dmnorm(muv[1:C], Sigmav.inv[1:C,1:C])

        for (c in 1:C) {
          theta0[i,c] = X[i,1:P]%*%beta[1:P,c]
        }

        Sigmae.inv[i,1:C,1:C] = inverse(Sigmae[i,1:C,1:C])
      }

  ## Priors:
  for (k in 1:p){
    for (c in 1:C){
      beta[k,c] ~ dnorm(0, 1/100)
    }
  }
  Sigmav.inv ~ dwish(Kv, 3)
  Sigmav = inverse(Sigmav.inv)
}
```

Simulation study: JAGS EB model specification

```
model{
      for(i in 1:m){
        y[i,1:C] ~ dmnorm(theta[i,1:C],  Sigmae.inv[i,1:C,1:C])
        theta[i,1:C] = theta0[i,1:C] + v[i,1:C]
        v[i,1:C] ~  dmnorm(muv[1:C], Sigmav.inv[1:C,1:C])

        for (c in 1:C) {
          theta0[i,c] =  x[i,1]*beta[c] + x[i,2:p]%*%betaF[2:p,c]
        }

        Sigmae.inv[i,1:C,1:C] = inverse(Sigmae[i,1:C,1:C])
      }

  for (c in 1:C){
        beta[c] ~ dnorm(0, 1/100)
      }

  Sigmav.inv ~ dwish(Kv, 3)
  Sigmav = inverse(Sigmav.inv)
}
```

Simulation study: STAN HB model specification

```
data {
    int<lower=0> m;
    int<lower=0> p;
    real x[m,p] ;
    vector[2] y[m] ;
    cov_matrix[2] Sigmae[m];
    cov_matrix[2] Kv;
}
parameters {
    real beta[p,2];
    vector[2] v[m] ;
    cov_matrix[2] Sigmav;
}
transformed parameters {
    vector[2] theta[m];
    for (i in 1:m) {
        vector[2] theta0;
        for (k in 1:2) {
           theta10[k] = 0;
           for (j in 1:p)
               theta0[k] += beta[j,k] * x[i,j];
        }
                theta[i] = v[i] + theta0;
    }
}

model {
    for (i in 1:p)
      for (k in 1:2)
         beta[i,k] ~ normal(0, 100);

    Sigmav ~ inv_wishart(3, Kv);

    for (i in 1:m)
       v[i] ~ multi_normal([0, 0], Sigmav);
    for (i in 1:m)
       y[i] ~ multi_normal(theta[i], Sigmae[i]);
}
```

Simulation study: STAN EB model specification

```
data {
    int<lower=0> m;
    int<lower=0> p;
    real x[m,p] ;
    vector[2] y[m] ;
    cov_matrix[2] Sigmae[m];
    cov_matrix[2] Kv;
    real betaF[p,2];
}
parameters {
    real beta[2];
    vector[2] v[m] ;
    cov_matrix[2] Sigmav;
}
transformed parameters {
    vector[2] theta[m];
    for (i in 1:m) {
        vector[2] theta0;
        for (k in 1:2) {
          theta0[k] <- beta[k] * x[i,1];
          for (j in 2:p)
              theta0[k] <- theta0[k] + betaF[j,k] * x[i,j];
        }
               theta[i] <- v[i] + theta0;
  }
}

model {
    for (k in 1:2)
          beta[k] ~ normal(0, 100);

    Sigmav ~ inv_wishart(3, Kv);

    for (i in 1:m)
        v[i] ~ multi_normal([0, 0], Sigmav);
    for (i in 1:m)
        y[i] ~ multi_normal(theta[i], Sigmae[i]);
}
```

Simulation study: Computational time results

Table 2: Computational Time Summaries

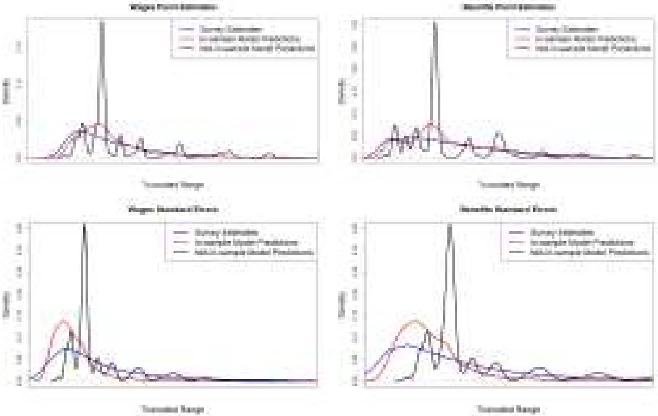
Model	Time (seconds)	
	HB	EB
S1	525	284
J1	5	2
J1l	35	11
S2	32744	14702
J2	1438	24
J2l	14227	109
S3	354416	76992
J3	18210	273
J3l	Too long ...	

Simulation study: Convergence results

Table 3: Convergence Results Summaries HB

Model	Approach	$\hat{R}$ (min, median, max)	MC Effective Sample Size (min, median, max)
S1	HB	(0.995, 1.000, 1.072)	(44, 571, 825)
	EB	(0.995, 1.005, 1.086)	(35, 342, 810)
J1	HB	(0.999, 1.004, 1.070)	(44, 600, 600)
	EB	(0.999, 1.023, 1.216)	(14, 440, 600)
J1l	HB	(1.001, 1.001, 1.030)	(130, 5800, 6000)
	EB	(1.001, 1.003, 1.014)	(190, 3400, 7500)
S2	HB	(0.995, 0.999, 1.289)	(14, 580, 908)
	EB	(0.995, 0.999, 1.262)	(9, 583, 802)
J2	HB	(0.999, 1.010, 1.818)	(6, 600, 600)
	EB	(1.037, 1.116, 2.029)	(5, 79, 600)
J2l	HB	(1.001, 1.002, 1.140)	(33, 6000, 6000)
	EB	(1.001, 1.002, 1.138)	(59, 3500, 7500)
S3	HB	(0.995, 1.000, 4.097)	(2, 582, 989)
	EB	(0.995, 1.003, 2.427)	(3, 503, 1109)
J3	HB	(0.999, 1.008, 2.118)	(5, 600, 600)
	EB	(0.999, 1.043, 3.003)	(1, 600, 600)

Example output from the small domain hierarchical Bayes multivariate model for employee compensation components*



- J1/2/3-type model ($m = 16,107$; $p = 18$)
- 16,107 survey estimates and in-sample model predictions, and 556,221 not-in-sample model predictions

* Erciulescu and Opsomer, 2019

Final thoughts...

- a flavor of innovation in official statistics programs
- existing tools and potential for development of novel ones
- software: more than just R JAGS and R STAN
- alternative sampling methods
- practical challenges: time + number of MCMC samples, large number of domains (in-sample and not-in-sample), variable selection, prior distributions for the linking model variance-covariance components, autocorrelation, cross-validation, visualization, storage
- beyond small area estimation - for example, bridging models for the Association of Fish and Wildlife Agencies (Erciulescu et al. 2019b)

Selected references

Bell, W. R., Basel W. W. , and Maples J. J. (2016). "An Overview of the U.S. Census Bureau's Small Area Income and Poverty Estimates Program." Analysis of Poverty Data by Small Area Estimation, M. Pratesi (Ed.). West Sussex: Wiley Sons, Inc., 349-77.

Casas Cordero Valencia, C., Encina J., and Lahiri P. (2016). "Poverty Mapping in Chilean Comunas." Analysis of Poverty Data by Small Area Estimation, M. Pratesi (Ed.). West Sussex: Wiley Sons, Inc., 379-403.

Elbers, C., Lanjouw, J. and Lanjouw, P. (2003). "Micro-Level Estimation of Poverty and Inequality." *Econometrica*, 71, 1, 355-364. http://www.jstor.org/stable/3082050.

Erciulescu, A. L. (2018). "Transparent and Reproducible Research in Agricultural Official Statistics." Government Advances in Statistical Programming Workshop. http://washstat.org/presentations/20181024/Erciulescu.pdf.

Erciulescu, A. L., Berg E. J., Cecere W., and Ghosh, M. (2018). "A Bivariate Hierarchical Bayesian Model for Estimating Cropland Cash Rental Rates at the County Level." *Survey Methodology, Statistics Canada, Catalogue No. 12-001-X*, 45, 2. https://www150.statcan.gc.ca/n1/pub/12-001-x/2019002/article/00001-eng.htm.

Erciulescu, A. L., Cruze N. B., and Nandram, B. (2019a). "Model-Based County Level Crop Estimates Incorporating Auxiliary Sources of Information." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182, 1, 283-303. doi:10.1111/rssa.12390.

Erciulescu, A.L., and Opsomer, J.D. (2019). "A model-based approach to predict employee compensation components." Joint Statistical Meetings, Denver, CO, July 2019. https://ww2.amstat.org/meetings/jsm/2019/onlineprogram/AbstractDetails.cfm?abstractid=300325.

Erciulescu, A.L., Opsomer, J.D., Breidt, J., and Morganstein, D. (2019b). "The data science challenge of bridging two complex surveys." New England Statistics Symposium.

Fay, 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). Taylor Francis: 269-77. doi:10.1080/01621459.1979.10482505.

Krenzke T., Mohadjer L., Li J., Erciulescu A. L., Fay R., Ren W., Van de Kreckhove W., Li L., Rao J. N. K. (forthcoming) "Program for the International Assessment of Adult Competencies. State and County Indirect Estimation Methodology." United States Department of Education. NCES2019012.

Lewandowski, D., Kurowicka, D., Joe, H. (2009). "Generating random correlation matrices based on vines and extended onion method." *Journal of Multivariate Analysis*, 100, 9, 1989-2001. doi: 10.1016/j.jmva.2009.04.008.

Thank you!

AndreeaErciulescu@westat.com