Small Area Estimation in the Annual Survey of Public Employment and Payroll (U.S. Census Bureau[1])

FCSM 2018, Washington D.C.
Bac Tran
Public Sector Sample Design & Estimation Branch, Chief
Economic Statistics Methods Division
U.S. Census Bureau

[1]*Disclaimer: Any views expressed are those of the author and not necessarily those of the U.S. Census Bureau*

Government Surveys

Annual Survey of Public Employment and Payroll (**ASPEP**)

- *Statistics on the number of federal, state, and local government employees and their gross payrolls*

Outline

- ❑ Target Population
- ❑ Parameters
- ❑ Sample Design
- ❑ Small Area Challenge
- ❑ Estimators
- ❑ Robust Estimation
- ❑ Evaluation
- ❑ Conclusions

Governmental Units

- *A government is an organized entity which, in addition to having governmental character, has sufficient discretion in the management of its own affairs to distinguish it as separate from the administrative structure of any other governmental unit.*
- Types of governmental unit
  - State Government (0)
  - Counties (1)
  - Cities (2)
  - Townships (3)
  - Special Districts (4)
  - School Districts (5)
  - Federal Government (6)

  ← Local Governments
- **Target Population**=Individual governmental units (about 92,000^{+} units)-The Governments Integrated Directory (GID)

Government Surveys (Cont'd)

ASPEP consists of three components

- Census of select federal agencies
- Census of 50 state governments
- A sample of about 10,000 + local governments units
  - Sampling frame is the most recent Census of Governments (CoG 2012)

Parameters of Interest

- Totals by
  - Type of government and functions
  - Level of government: Local, state, and state and local
  - **ASPEP** estimates (2015)

itemname	stateab	ftemp15	ftpay15	totpay15	totemp15
Total	US	14,439,303	69,001,541,148	75,070,674,087	19,334,121
Education	US	7,562,046	34,455,692,379	38,577,353,203	11,041,338
Higher education	US	1,559,218	8,884,420,759	10,909,241,441	3,265,391
Instructional employees	US	512,484	3,949,633,303	5,039,544,144	1,147,101
Other employees	US	1,046,734	4,934,787,456	5,869,697,297	2,118,290
Elementary and secondary educ	US	5,922,868	25,193,849,851	27,268,214,821	7,683,263
Instructional employees	US	4,265,938	20,242,582,206	21,444,188,536	5,241,518
Other employees	US	1,656,930	4,951,267,645	5,824,026,285	2,441,745
Other education	US	79,960	377,421,769	399,896,941	92,684
Libraries	US	86,970	345,998,733	449,082,651	184,206

Source: U.S. Census Bureau, 2015 Annual Survey of Employment and Payroll

Sample Design

Multistage Sample Design

- Proportional-To-Size (πPS) sample
  → Stratified πPS (state x type) based on Total Pay (MoS)

- Sample units are grouped into two strata depending on their sizes and then subsample in the stratum with small units (modified cutoff sampling)

Small Area Challenges

- Designed at (state, type) level, estimated at function levels
- Structural zeros & small sampling rates
  →cells in which observations are impossible
- Outliers

Estimators

- Direct Estimator (Horvitz-Thompson)
- Calibration
- Generalized Regression
- Synthetic
- Ratio
- Structure PREserving Estimation (SPREE)
- Composite
- Batesse-Harter-Fuller (BHF) Model
- Mixed Models (EBLUP)
- Mixture Models (Design-based/Bayesian)

Mixed Models (Empirical Best Linear Unbiased Predictor)-With Data Transformation

❑ Mixed Model (area-level covariate)

$$\log(y_{ij}) = \beta_0 + \beta_1 \log(\bar{X}_i) + v_i + e_{ij}$$

$$v_i \overset{iid}{\sim} N(0, \tau^2) \text{ and } e_{ij} \overset{iid}{\sim} N(0, \sigma^2) \text{ where } i = \text{area } i^{th} \text{ and } j = \text{unit } j^{th}$$

❑ Mixed Model (unit-level covariate)

$$\log(y_{ij}) = \beta_0 + \beta_1 \log(X_{ij}) + v_i + e_{ij}$$

$Two - level\ Model$

1. $\log(y_{ij}) \mid \theta_i \sim N(0, \sigma^2)$
2. $\theta_i \mid \beta, \tau \sim N(X_i'\beta, \tau)$

Data (ASPEP)

California 2007 & 2012 Census

Government units that overlap

between the 2007 and 2012 Census

of Government units reporting strictly positive numbers of full-time employees.

Design-Based Simulation ASPEP

- Universe

  Government units that overlap between the 2007/2012 Censuses of Governments reporting strictly positive numbers of full-time employees. For simplicity, use California data.

- Simulation

  →Replicate 1000 samples containing 29 small areas (functions) using production sample design from the universe

  →Apply different estimation methods on each sample

- Variable of Interest

  →The number of full-time employees/finance functions

Evaluation Measures

- Compare the performances of the estimators using Relative Bias (RB) and Relative Root Mean Squared Error (RRMSE) over 1000 sample replicates

- $RB_i = \frac{1}{1000}\sum_{r=1}^{1000}\frac{(\hat{y}_i^{(r)} - Y_i)}{Y_i}$

- $RRMSE_i = \sqrt{\frac{1}{1000}\sum_{r=1}^{1000}(\hat{y}_i^{(r)} - Y_i)^2}\,/Y_i$

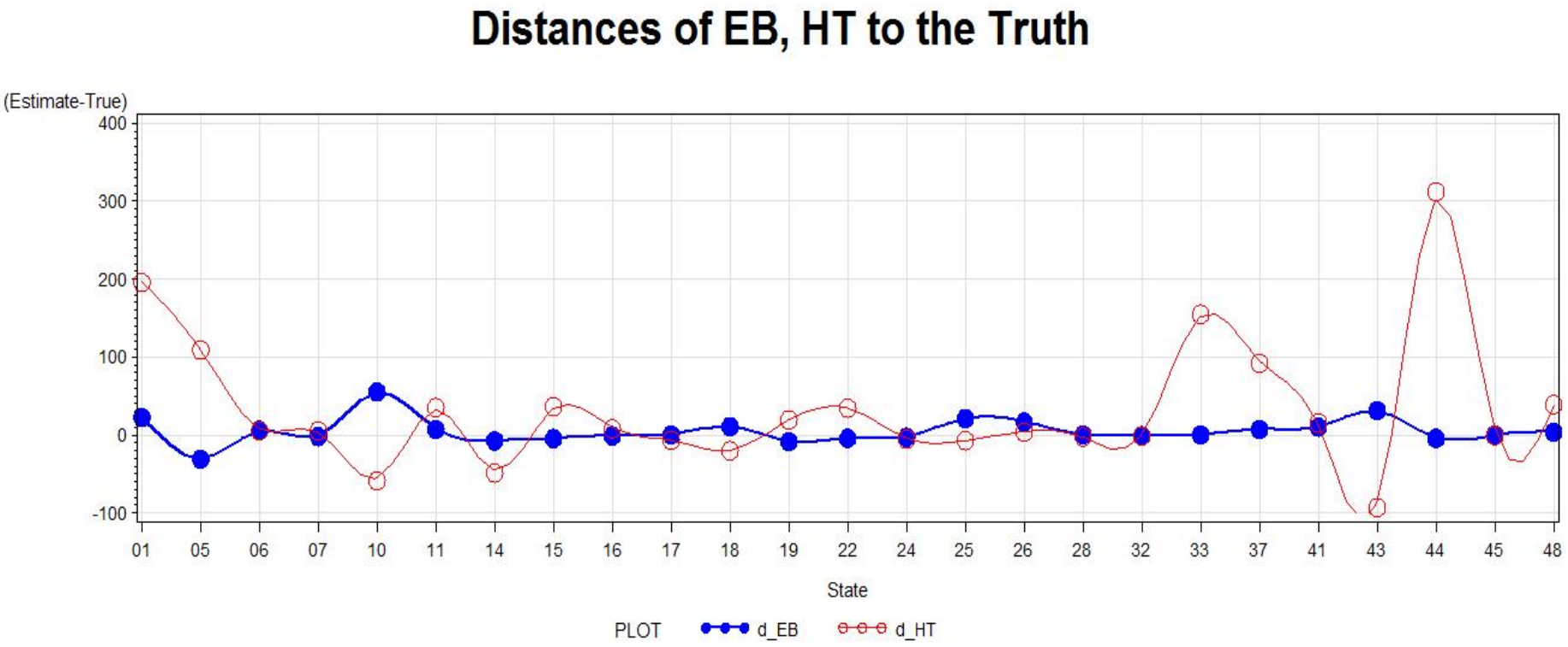
Result 1 (cont'd)

Table 1: Estimator Performances (California State)

$\widehat{mse}$					$\widehat{bias}$				
Number of times the model outperforms the others					*Number of times the model has less Biases vs the others*				
EB-Unit	EB-Area	Composite	HT	SPREE	EB-Unit	HT	EB-Area	Composite	SPREE
507	317	273	66	59	33	75	79	266	769

Result 1 (Cont'd)
(Gas Supply, All States, Average n= 4)

Figure 1



Conclusion 1

- Unit-level covariate EBLUP outperforms the Direct, Synthetic, Composite, BHF, Area-level covariate estimators in small areas

- Direct estimator is reliable in big areas

Outlier Challenges

- Robust Estimation
- Mixture Model (Design-based)
  (Gershunskaya and Lahiri 2010/2011 papers)
- Mixture Model (hierarchical Bayes)

Robust Estimation

Hierarchical Bayes with t-distribution and type of the government unit as fixed effect

$\boldsymbol{Model}$: $y_{ijk} = \beta_0 + \beta_1 x_{ijk} + \alpha_j + u_i + e_{ijk}$

y_{ijk}, x_{ijk} *are the number of full time employees from survey data and census year in* log *scale , respectively, where* $i =$ *area* i *(function code),* $j =$ *type of government,* $k =$ *unit* k^{th}, $i = 1,2,..,29$ *areas,* $j = 1,2,3,4$ *type of government,* $k = 1,2,..,N_i$

α_j *is fixed effect of government type* j^{th}

u_i *is random effect of function code* i^{th}

$e_{ijk} | \sigma_e^2 \overset{iid}{\sim} N(0, \sigma_e^2)$

$u_i | \sigma_u^2 \overset{iid}{\sim} N(0, \sigma_u^2)$

Prior

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} const & 0 \\ 0 & const \end{bmatrix}\right)$$

$\sigma_u^2 \sim inverse\ gamma(0.01, 0.01)$

$\sigma_e^2 \sim inverse\ gamma(0.01, 0.01)$

$flat\ prior\ on\ \alpha_j$

Likelihood

$y_{ijk} | u_i, \alpha_j \sim t(mean = \beta_0 + \beta_1 x_{ijk} + \alpha_j + u_i, \sigma_e^2, df = 4)$

Mixtures of Two Normal Distributions (Design-based)

$$Model:\ y_{ij} = \boldsymbol{X}_{ij}^{T}\beta + u_i + e_{ij}$$

$$where\ i = area\ i^{th}\ and\ j = unit\ j^{th} of\ area\ i^{\text{th}}$$

$$u_i \overset{iid}{\sim} N(0, \tau^2)$$

$$e_{ij} | z \overset{iid}{\sim} (1-z)N(0, \sigma_1^2) + z\, N(0, \sigma_2^2), \sigma_2 > \sigma_1$$

$$z \mid \pi \sim Bin(1, \pi)$$

$$The\ parameter\ \theta = (\sigma_1, \sigma_2, \tau, \pi, \beta) is\ estimated\ by\ an\ EM\ algorithm$$

$$(See\ Giang\ \&\ Tran\ JSM\ 2016)$$

Mixtures of Two Normal Distributions (Hierarchical Bayes)

Model: $y_{ij} = x_{ij}^T \beta + u_i + e_{ij}$

$u_i|\sigma_u^2 \overset{iid}{\sim} N(0, \sigma_u^2)$

$z|\pi \sim Bin(1, \pi)$

$\pi = \frac{1}{(1+e^{-z})}$

$e_{ij}|z, \sigma_1^2, \sigma_2^2, z \overset{iid}{\sim} (1-z)N(0, \sigma_1^2) + zN(0, \sigma_2^2)$

$\sigma_u^2 \sim inverse\ gamma$

$\sigma_1^2 \sim inverse\ gamma$

$\sigma_2^2 \sim inverse\ gamma$

$$\begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix} \sim N\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}50 & 0 \\ 0 & 50\end{bmatrix}\right)$$

$\boldsymbol{Likelihood}$: $y_{ij}|\sigma_1^2,\ \sigma_2^2,\ u_i \sim \pi * N(0, \sigma_1^2) + (1-\pi) * N(0, \sigma_2^2)$

$(See\ Giang\ \&\ Tran\ JSM\ 2017)$

Simulation (ASPEP)

- Generate 1000 samples using 2014 production sample design
- For each sample replicate, run 12,500 MCMC iterations, burn-in samples of 2500, skip interval 5 → 2500 MCMC final samples
- Using the hierarchical model, predict the total number of employees for all units not in the sample
- For each area, sum the total number of employees for all sampled units and predicted total number of employees for all non-sampled units to obtain the final estimate

Result 2-ASPEP

Table 2a: Relative Root Mean Squared Errors (RRMSE) of Six Different Estimators (California Data)

Function	EBLUP	EBFixedType	HB	HBFixedType	N2Design	HBMixture
001	0.52%	0.44%	0.52%	0.61%	1.03%	0.29%
005	0.67%	0.64%	0.52%	0.48%	0.57%	0.57%
012	1.30%	1.67%	1.29%	0.98%	1.30%	1.65%
016	4.09%	4.13%	2.79%	2.23%	2.96%	2.53%
018	4.35%	3.19%	3.50%	2.76%	3.43%	2.50%
023	0.84%	0.81%	0.68%	0.68%	2.27%	0.97%
024	0.83%	2.80%	0.72%	0.55%	2.92%	1.20%
025	0.40%	0.47%	0.42%	0.46%	0.58%	0.58%
029	1.20%	1.66%	1.14%	1.07%	1.11%	1.49%
032	0.83%	0.69%	0.69%	0.57%	0.58%	0.35%
040	0.63%	0.83%	0.42%	0.43%	2.55%	0.78%
044	1.69%	1.09%	0.94%	0.91%	1.93%	0.42%
050	3.50%	0.94%	1.94%	1.37%	2.85%	0.69%
052	0.80%	0.51%	0.64%	0.62%	0.79%	0.93%
059	7.91%	2.51%	3.41%	1.98%	5.32%	3.92%
061	3.00%	0.81%	1.71%	1.54%	0.92%	1.55%
062	0.66%	2.71%	0.33%	0.45%	1.15%	0.39%
079	0.68%	0.72%	0.29%	0.28%	0.19%	0.25%
080	1.81%	2.02%	1.13%	1.36%	7.29%	1.25%

Result 2- ASPEP (Cont'd)

Table 2b: Relative Root Mean Squared Errors (RRMSE) of Six Different

Function	EBLUP	EBFixedType	HB	HBFixedType	N2Design	HBMixture
081	2.48%	1.67%	1.83%	1.73%	1.67%	1.35%
087	1.35%	1.39%	1.19%	1.03%	2.18%	1.18%
089	2.09%	2.33%	2.67%	2.36%	0.98%	1.98%
091	2.19%	2.96%	1.25%	0.95%	4.44%	0.38%
092	0.46%	0.60%	0.43%	0.45%	3.36%	0.51%
093	1.54%	1.65%	1.70%	1.76%	4.16%	1.80%
094	1.19%	1.19%	0.92%	0.86%	0.56%	1.07%
112	1.50%	1.17%	1.36%	0.99%	0.57%	0.23%
124	1.49%	4.15%	2.03%	2.43%	5.98%	2.74%
162	1.26%	1.31%	0.68%	0.73%	1.34%	0.50%
	3	2	5	7	3	9

Conclusions 2

- EBLUP outperforms the other estimators Direct/Synthetic/Ratio/SPREE/Composite/BHF
- Mixture models (design-based) outperforms EBLUP, especially in cells where outliers exits
- The same model but hierarchical Bayes outperforms design-based mixture model
- Hierarchical Bayes is favorable because HB is flexible in terms of handling transformation, benchmarking, measure of uncertainty and interval estimation
- 'Hybrid' Estimators

Future Research

- Mixture of two distributions other than normal distributions
- Mixture of normal distributions with more than two mixing components
- Correlation across areas
- Applying the methods to other surveys

Appendix
Some Function Codes (ASPEP)

Function	Descriptions
000	Totals for Government
001	Air Transportation
002	Space Research & Technology (Federal)
005	Correction
006	Nat Defense & International Relations (Federal)
012	Elementary and Secondary - Instruction
014	Postal Service (Fed)
016	Higher Education - Other
018	Higher Education - Instructional
021	Other Education (State)
022	Social Insurance Administration (State)
023	Financial Administration
024	Firefighters
025	Judicial & Legal
029	Other Government Administration
032	Health
040	Hospitals
044	Highways
050	Housing & Community Development

Appendix
Some Function Codes (ASPEP)

052	Libraries
059	Natural Resources
061	Parks & Recreation
062	Police Protection - Officers
079	Public Welfare
080	Sewerage
081	Solid Waste Management
087	Water Transport & Terminals
089	All Other & Unallocable
090	Liquor Stores (State)
091	Water Supply
092	Electric Power
093	Gas Supply
094	Transit
112	Elementary and Secondary - Other
124	Fire - Other
162	Police-Other

Question?

Thank you for your attendance

Contact

Bac.tran@census.gov