IAOS-ISI 2024
MEXICO CONFERENCE

Modern Approaches for Small Area Estimation in Official Statistics

Instructions: Click on the link to access each author's presentation.

Organiser: Paul Parker

Participants:

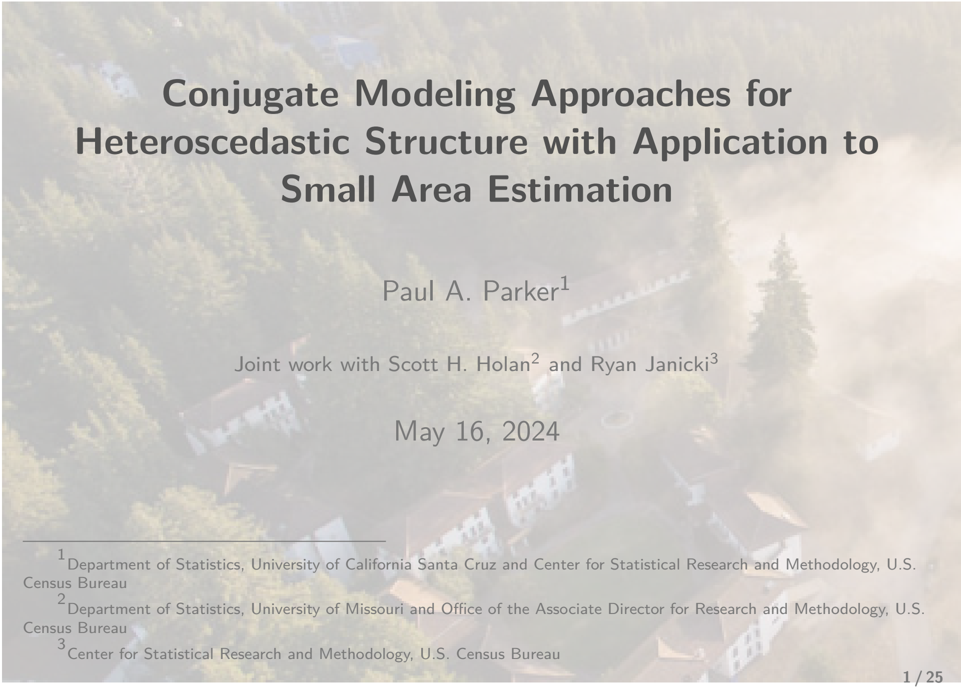
Paul Parker: Conjugate Modeling Approaches for Heteroscedastic Structure with Application to Small Area Estimation

Gauri Datta: A credible region for the rank vector of many subpopulations

Benjamin Schneider: Statistical data integration using multilevel models to predict employee compensation

Scott H. Holan:* Bayesian Unit-level Models for Longitudinal Survey Data under Informative Sampling

* Work presentation not available or non-existent



Conjugate Modeling Approaches for Heteroscedastic Structure with Application to Small Area Estimation

Paul A. Parker[1]

Joint work with Scott H. Holan[2] and Ryan Janicki[3]

May 16, 2024

[1] Department of Statistics, University of California Santa Cruz and Center for Statistical Research and Methodology, U.S. Census Bureau

[2] Department of Statistics, University of Missouri and Office of the Associate Director for Research and Methodology, U.S. Census Bureau

[3] Center for Statistical Research and Methodology, U.S. Census Bureau

Disclaimer

This work is to inform interested parties of ongoing research and to encourage discussion. The views expressed on statistical issues are those of the authors and not those of the U.S. Census Bureau.

Background

Small Area Estimation

Goal: Estimate a population characteristic of interest aggregated over geographic areas (or other domains) based on survey sample data.

Area-Level Models

- Treat the direct estimate as the response variable
- Usually incorporate smoothing through the model
- May be the only option for analysts outside of a statistical agency

Unit-Level Models

- Treat individual survey unit responses as response variables
- Can make predictions for all units in the population
- Typically feasible within a statistical agency

Area-Level Models

Fay III and Herriot (1979)

$$y_i|\theta_i, \sigma_i^2 \overset{ind}{\sim} \mathrm{N}(\theta_i, \sigma_i^2), \; i = 1, \ldots, d$$
$$\theta_i = \boldsymbol{x}_i{}'\beta + \eta_i$$
$$\eta_i \overset{iid}{\sim} \mathrm{N}(0, \sigma_\eta^2),$$

- y_i is the direct estimate of a population quantity for area i
- θ_i is the latent population quantity of interest
- σ_i^2 is the design-based variance of y_i
- $\boldsymbol{x}_i$ is a vector of covariates for area i

Note:

The design-based variance, σ_i^2, is assumed to be known, but in practice, $s_i^2 = \hat{\sigma}_i^2$ estimated from the data and plugged in

Modeling Unknown Sampling Variance

You and Chapman (2006)

$$
\begin{aligned}
y_i|\theta_i, \sigma_i^2 &\overset{ind}{\sim} \mathrm{N}(\theta_i, \sigma_i^2), \;\; i = 1, \ldots, d \\
s_i^2|\sigma_i^2 &\overset{ind}{\sim} \mathrm{Gamma}\left(\frac{n_i - 1}{2}, \frac{n_i - 1}{2\sigma_i^2}\right), \;\; i = 1, \ldots, d \\
\theta_i &= \mathbf{x}_i{}'\beta + \eta_i \\
\eta_i &\overset{iid}{\sim} \mathrm{N}(0, \sigma_\eta^2) \\
\sigma_i^2 &\overset{ind}{\sim} \mathrm{IG}(a_i, b_i).
\end{aligned}
$$

- A data model for s_i^2 is introduced, conditional on the true but unknown σ_i^2
- n_i represents the sample size in area i

Note:

The data model for s_i^2 assumes a simple random sample within area i. More careful consideration may be necessary for complex sample designs.

Modeling Unknown Sampling Variance

Sugasawa et al. (2017)

A Bayesian extension considers covariates in the variance model:

$$
\begin{aligned}
y_i|\theta_i, \sigma_i^2 &\overset{ind}{\sim} \mathrm{N}(\theta_i, \sigma_i^2), \ i = 1, \ldots, d \\
s_i^2|\sigma_i^2 &\overset{ind}{\sim} \mathrm{Gamma}\left(\frac{n_i - 1}{2}, \frac{n_i - 1}{2\sigma_i^2}\right), \ i = 1, \ldots, d \\
\theta_i &= \boldsymbol{x}_i{}'\beta_1 + \eta_i \\
\eta_i &\overset{iid}{\sim} \mathrm{N}(0, \sigma_\eta^2) \\
\sigma_i^2 &\overset{ind}{\sim} \mathrm{IG}\left(a_i, b_i \exp(\boldsymbol{x}_i{}'\beta_2)\right).
\end{aligned}
$$

Note:

The typical Gaussian prior for β_2 requires a Metropolis-Hastings sampler, which can be very difficult to tune, especially in high dimensions.

Methodology

The Multivariate Log-Gamma Distribution

The cornerstone of our modeling framework is the Multivariate Log-Gamma distribution (MLG).

- Introduced by Bradley et al. (2018) and Bradley et al. (2019)
  - Used to Model dependent data using a Poisson likelihood
- Used by Parker et al. (2021) to model heteroskedastic data via a negative log link function for the variance
  - Acts as a conjugate prior for variance regression parameters

Methodology

MLG Density

$$f(\boldsymbol{y}) = \det(\boldsymbol{V}^{-1}) \left\{ \prod_{i=1}^{n} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)} \right\} \exp\left[\boldsymbol{\alpha}' \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) - \boldsymbol{\kappa}' \exp\left\{ \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \right\} \right], \tag{1}$$

- Denoted by MLG$(\boldsymbol{\mu}, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$
- The length n vector $\boldsymbol{\mu}$ acts as a centrality parameter
- The $n \times n$ matrix $\boldsymbol{V}$ controls the correlation structure
- The length n vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ are shape and rate parameters respectively

Methodology

MLG Relation to the Normal Distribution

Another important result given by Bradley et al. (2018) is that $\text{MLG}(\mathbf{c}, \alpha^{1/2}\mathbf{V}, \alpha\mathbf{1}, \alpha\mathbf{1})$ converges in distribution to a multivariate normal distribution with mean $\mathbf{c}$ and covariance matrix $\mathbf{VV}'$ as the value of α approaches infinity.

This allows for the use of MLG priors in place of Gaussian priors, in situations where it is computationally preferable, while still achieving the same prior in the limit of α.

Proposed Approach

Heteroskedastic Area-Level Model (HALM)

$$
\begin{aligned}
y_i|\theta_i,\sigma_i^2 &\overset{ind}{\sim} \mathrm{N}(\theta_i,\sigma_i^2),\ i=1,\ldots,d\\
s_i^2|\sigma_i^2 &\overset{ind}{\sim} \mathrm{Gamma}\left(\frac{n_i-1}{2},\frac{n_i-1}{2\sigma_i^2}\right),\ i=1,\ldots,d\\
\theta_i &= \boldsymbol{x}_i{}'\boldsymbol{\beta}_1+\eta_{1i}\\
-\log(\sigma_i^2) &= \boldsymbol{x}_i{}'\boldsymbol{\beta}_2+\eta_{2i}\\
\boldsymbol{\eta}_1|\sigma_{\eta_1}^2 &\sim \mathrm{N}(\mathbf{0},\sigma_{\eta_1}^2\mathbf{I})\\
\boldsymbol{\eta}_2|\sigma_{\eta 2}^2 &\sim \mathrm{MLG}(\mathbf{0},\alpha^{1/2}\sigma_{\eta_2}\mathbf{I},\alpha\mathbf{1},\alpha\mathbf{1})\\
\boldsymbol{\beta}_1 &\sim \mathrm{N}(\mathbf{0},\sigma_\beta^2\mathbf{I})\\
\boldsymbol{\beta}_2 &\sim \mathrm{MLG}(\mathbf{0},\alpha^{1/2}\sigma_\beta\mathbf{I},\alpha\mathbf{1},\alpha\mathbf{1})
\end{aligned}
$$

Proposed Approach

Spatial Heteroskedastic Area-Level Model (SHALM)

The HALM can be modified to allow spatial dependence structure by considering spatially correlated prior structures for the random effects:

$$\boldsymbol{\eta}_1|\sigma^2_{\eta_1} \sim \text{N}\left(\mathbf{0}, \sigma^2_{\eta_1}(\boldsymbol{D}-\boldsymbol{W})^{-1}\right)$$
$$\boldsymbol{\eta}_2|\sigma^2_{\eta 2} \sim \text{MLG}(\mathbf{0}, \alpha^{1/2}\sigma_{\eta_2}(\boldsymbol{D}-\boldsymbol{W})^{-1/2}, \alpha\mathbf{1}, \alpha\mathbf{1})$$

This prior for $\boldsymbol{\eta}_1$ follows an intrinsic conditional auto-regressive (ICAR) structure (Besag et al., 1991). The prior for $\boldsymbol{\eta}_2$ is asymptotically equivalent to an ICAR prior.

Note:

Here, $\boldsymbol{W}$ represents an area adjacency matrix and $\boldsymbol{D}$ is a diagonal matrix with entry D_{ii} corresponding to the number of neighbors shared with area i

Unit-Level Models

Battese et al. (1988)

$$
\begin{aligned}
y_{ij} &\overset{ind}{\sim} \mathrm{N}(\mu_{ij}, \sigma^2) \\
\mu_{ij} &= \boldsymbol{x}_{ij}'\beta + \eta_i \\
\eta_i &\overset{iid}{\sim} \mathrm{N}(0, \sigma_\eta^2)
\end{aligned}
$$

- y_{ij} is the sample response for unit j in area i
- $\boldsymbol{x}_{ij}$ is a vector of unit-level covariates
- η_i is a random effect for area i

Note:

This model assumes that the survey design is ignorable. Bias will be introduced in the case of an informative sample design.

Bayesian Pseudo-likelihood

A pseudo-likelihood (PL) approach may be used in a Bayesian setting to account for informative sampling. Savitsky and Toth (2016) show that the use of a PL within a Bayesian model results in a pseudo-posterior distribution that converges to the population generating distribution

$$\hat{\pi}(\boldsymbol{\theta}|\mathbf{y}, \tilde{\mathbf{w}}) \propto \left\{ \prod_{j \in \mathcal{S}} f(y_j|\boldsymbol{\theta})^{\tilde{w}_j} \right\} \pi(\boldsymbol{\theta}).$$

In this case, $\tilde{w}_j$ represents the survey weights after scaling to sum to the sample size.

Predictions can then be made for the entire population via the posterior predictive distribution, and aggregated as necessary to create population estimates.

Proposed Approach

Heteroskedastic Unit-Level Model (HULM)

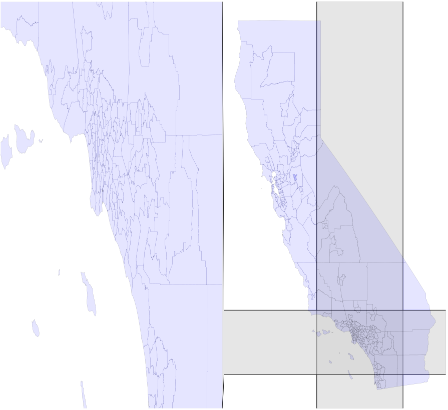
$$
\begin{aligned}
\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\sigma}^{2} &\propto \prod_{j\in\mathcal{S}} \mathrm{N}(y_{ij}|\mu_{ij},\sigma_{ij}^{2})^{\tilde{w}_{ij}} \\
\mu_{ij} &= \boldsymbol{x}_{ij}'\boldsymbol{\beta}_1 + \eta_{1i} \\
-\log(\sigma_{ij}^{2}) &= \boldsymbol{x}_{ij}'\boldsymbol{\beta}_2 + \eta_{2i} \\
\boldsymbol{\eta}_1|\sigma_{\eta_1}^{2} &\sim \mathrm{N}(\mathbf{0},\sigma_{\eta_1}^{2}) \\
\boldsymbol{\eta}_2|\sigma_{\eta 2}^{2} &\sim \mathrm{MLG}(\mathbf{0},\alpha^{1/2}\sigma_{\eta_2}\mathbf{I},\alpha\mathbf{1},\alpha\mathbf{1}) \\
\boldsymbol{\beta}_1 &\sim \mathrm{N}(\mathbf{0},\sigma_{\beta}^{2}) \\
\boldsymbol{\beta}_2 &\sim \mathrm{MLG}(\mathbf{0},\alpha^{1/2}\sigma_{\beta}\mathbf{I},\alpha\mathbf{1},\alpha\mathbf{1})
\end{aligned}
$$

Empirical Simulation

Data

We work with the American Community Survey public use microdata sample (PUMS)

- Limited to 2018 1-year PUMS sample
- Subset to the state of California
  - Roughly 179,000 individuals
  - 265 public use microdata areas (PUMAs)



Empirical Simulation

Setup

We treat the existing PUMS data as a population for which we would like to estimate population quantities.

- We use two different sampling methods
  - Stratified random sampling by PUMA, taking a simple random sample without replacement of 5 observations per area
  - Taking a probability proportional to size sample with size variable constructed from the original survey weight as well as respondent income
- Using the sub-sampled data, we estimate the population average income by PUMA
- We fit all models after log transforming the response
- We consider the log population size as an area-level covariate and we consider age, sex, and race as unit-level covariates

Empirical Simulation

Models

- We consider two unit-level models
  - A weighted pseudo-likelihood version of the model used by Battese et al. (1988) (PL-BULM)
  - The proposed heteroskedastic unit-level model (HULM)

- We consider four area-level models
  - The model used by Fay III and Herriot (1979) (FH)
  - The model used by Sugasawa et al. (2017) (STK)
  - The proposed heteroskedastic area-level model (HALM)
  - The proposed spatial heteroscedastic area-level model (SHALM)

Note:

A Gaussian unit-level model for income is a starting point, but more complex methodology would be ideal in this case

Empirical Simulation

Metrics of Comparison

Root Mean Squared Error (RMSE)

$$\sqrt{\sum_{k=1}^{K} \frac{(\hat{\theta}_k - \theta)^2}{K}}$$

- Considers both bias and variance of the point estimates

Interval Score (Gneiting and Raftery, 2007)

$$\tfrac{1}{K}\sum_{k=1}^{K}\left\{(u_k-\ell_k)+\tfrac{2}{\alpha}(\ell_k-\theta)I(\theta<\ell_k)+\tfrac{2}{\alpha}(\theta-u_k)I(\theta>u_k)\right\}$$

- Considers both the width and the coverage rate of the interval estimate
- Evaluated for $\alpha = 0.05$ (i.e., a 95% credible interval)

Empirical Simulation

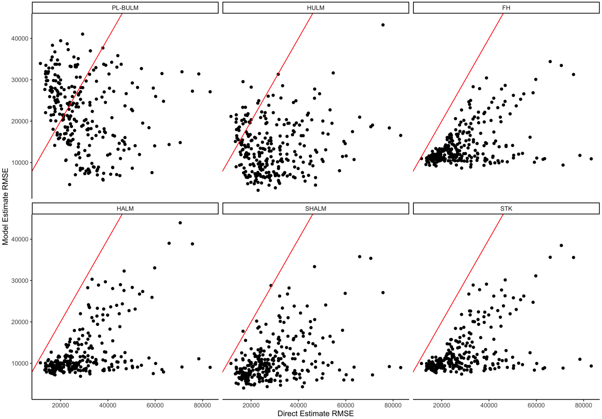
Stratified Sampling Summary

Estimator	Rel. RMSE	Abs. Bias ($\times 10^4$)	Cov. Rate	Int. Score ($\times 10^4$)
PL-BULM	1.080	20.299	0.368	30.570
HULM	0.687	11.356	0.596	14.720
FH	0.694	5.126	0.894	8.790
HALM	0.640	7.677	0.956	6.695
SHALM	0.561	6.759	0.933	6.031
STK	0.636	6.958	0.952	6.648

Table: Empirical simulation results for stratified random sampling by Public-Use Microdata Area

Empirical Simulation

Stratified Sampling RMSE by PUMA



Empirical Simulation

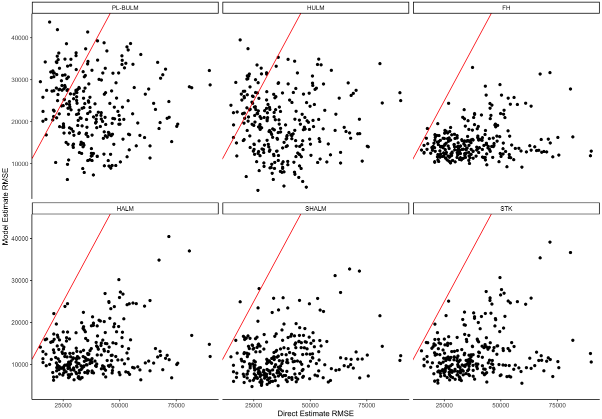
PPS Sampling Summary

Estimator	Rel. RMSE	Abs. Bias ($\times 10^4$)	Cov. Rate	Int. Score ($\times 10^4$)
PL-BULM	0.766	19.607	0.392	30.846
HULM	0.646	16.509	0.461	23.574
FH	0.492	6.815	0.955	8.855
HALM	0.442	9.765	0.958	6.800
SHALM	0.401	8.481	0.913	6.267
STK	0.429	9.195	0.958	6.468

Table: Empirical simulation results for probability proportional to size sampling

Empirical Simulation

PPS RMSE by PUMA



Summary

- We introduce an area-level model that uses covariates to shrink both the direct estimate means and variances
    - Use of the multivariate log-gamma distribution allows for computational efficiency
    - Computational efficiency allows for extension to spatial modeling for both the mean and variance
- We introduce a computationally efficient heteroskedastic data model for unit-level survey data under informative sampling
- We illustrate the proposed methodology via an empirical simulation study

Thank you!

paulparker@ucsc.edu

The paper can be found at the Journal of Survey Statistics and Methodology using the QR code below:

SCAN ME

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20.

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2019). Bayesian hierarchical models with conjugate full-conditional distributions for dependent data from the natural exponential family. *Journal of the American Statistical Association*, pages 1–16.

Bradley, J. R., Holan, S. H., Wikle, C. K., et al. (2018). Computationally efficient multivariate spatio-temporal models for high-dimensional count-valued data (with discussion). *Bayesian Analysis*, 13(1):253–310.

Fay III, 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):269–277.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Parker, P. A., Holan, S. H., and Wills, S. A. (2021). A general bayesian model for heteroskedastic data with fully conjugate full-conditional distributions. *Journal of Statistical Computation and Simulation*, 91(15):3207–3227.

Savitsky, T. D. and Toth, D. (2016). Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 10(1):1677–1708.

Sugasawa, S., Tamae, H., and Kubokawa, T. (2017). Bayesian estimators for small area models shrinking both means and variances. *Scandinavian Journal of Statistics*, 44(1):150–167.

You, Y. and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32(1):97.

Uses the Beamer `simple` theme from `http://github.com/famuvie/beamerthemesimple`

Credible Distributions for Ranking of Entities

Gauri S. Datta

Department of Statistics, University of Georgia
Center for Statistical Research & Methodology, U.S. Census Bureau

May 16, ISI-IAOS Meeting, Mexico City, May 15-17, 2024
In collaboration with Yiren Hou and Abhyuday Mandal

Disclaimer:

This presentation is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the presenter and not those of the U.S. Census Bureau.

Outline

1. Introduction
2. Review of KWW (2020) Framework
3. Proposed Bayesian Framework
4. Baseball Example – Efron and Morris (1975)
5. Commuting Times Example – Klein et al. (2020)
6. Median Incomes Example
7. Summary and Conclusions

Outline

1 Introduction

2 Review of KWW (2020) Framework

3 Proposed Bayesian Framework

4 Baseball Example – Efron and Morris (1975)

5 Commuting Times Example – Klein et al. (2020)

6 Median Incomes Example

7 Summary and Conclusions

Motivation

- Inference on overall ranking of a set of entities, such as chess players, subpopulations or hospitals, is an important problem.
- Accountability of public institutions involve in making quantitative comparisons between institutions in the areas of health and education.
- Inference of ranks based on point estimates of means does not account for the uncertainty in those estimates.
- Goldstein and Spiegelhalter (1996) and Klein, Wright and Wieczorek (2020) recognized treating estimated ranks without regard for uncertainty is problematic.
- KWW proposed a comprehensive frequentist solution. Following GS Bayesian approach, we propose a comprehensive Bayesian solution.

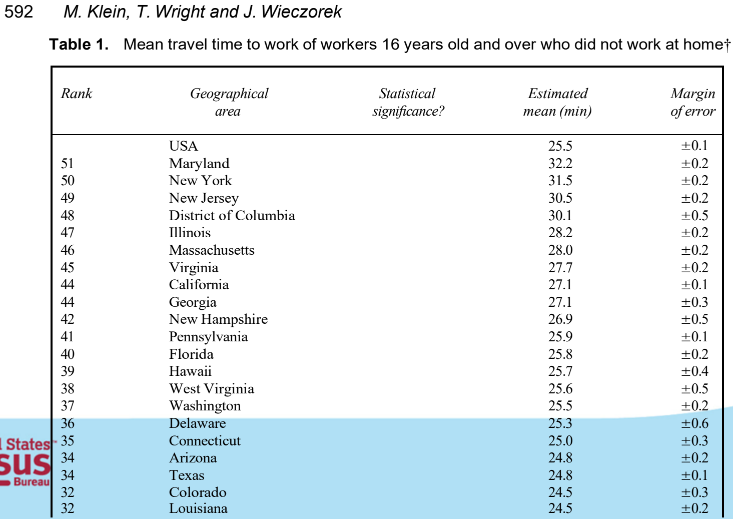
Main points of the talk

- Productions of accurate estimates of the means $\theta_1, \cdots, \theta_m$ for some characteristic for the subpopulations are usually the primary goals.
- It is also important to accurately identify subpopulations that are either at the upper or at the lower end in terms of their means. This goal requires accurately estimating the ranks of several or all the subpopulations.
- The importance of joint ranking of m subpopulations with unknown means of a common characteristic has been emphasized by KWW.

Importance of estimation of overall ranking

- KWW cautioned that from published point estimates of the means with no explicit ranking, practitioners frequently naively ascertain ranks of the subpopulations. Ranks determined this way are only point estimates, ignoring uncertainty.
- Even the best possible estimators of small area means are subject to error due to sampling variability. It is both imperative and a sound policy to evaluate uncertainty associated with the reported ranks based on reasonable estimators of means.
- KWW used the American Community Survey (ACS) to rank 50 U.S. states and DC, using mean commuting times of workers 16 years old and over, not working from home. They determined joint confidence set for the true rank vector from that of the true means.

Travel time data (upper half of Table 1 of KWW paper)

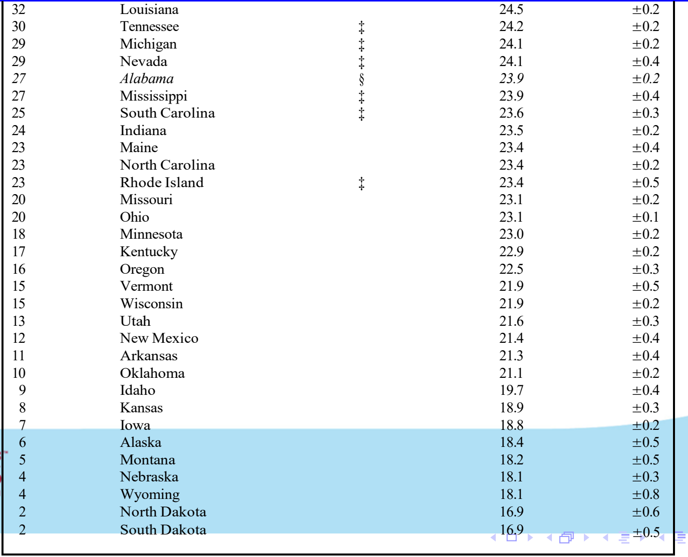


592 *M. Klein, T. Wright and J. Wieczorek*

Table 1. Mean travel time to work of workers 16 years old and over who did not work at home†

Rank	*Geographical area*	*Statistical significance?*	*Estimated mean (min)*	*Margin of error*
	USA		25.5	±0.1
51	Maryland		32.2	±0.2
50	New York		31.5	±0.2
49	New Jersey		30.5	±0.2
48	District of Columbia		30.1	±0.5
47	Illinois		28.2	±0.2
46	Massachusetts		28.0	±0.2
45	Virginia		27.7	±0.2
44	California		27.1	±0.1
44	Georgia		27.1	±0.3
42	New Hampshire		26.9	±0.5
41	Pennsylvania		25.9	±0.1
40	Florida		25.8	±0.2
39	Hawaii		25.7	±0.4
38	West Virginia		25.6	±0.5
37	Washington		25.5	±0.2
36	Delaware		25.3	±0.6
35	Connecticut		25.0	±0.3
34	Arizona		24.8	±0.2
34	Texas		24.8	±0.1
32	Colorado		24.5	±0.3
32	Louisiana		24.5	±0.2

Travel time data (lower half of Table 1 of KWW paper)



32	Louisiana		24.5	±0.2
30	Tennessee	‡	24.2	±0.2
29	Michigan	‡	24.1	±0.2
29	Nevada	‡	24.1	±0.4
27	*Alabama*	§	*23.9*	*±0.2*
27	Mississippi	‡	23.9	±0.4
25	South Carolina	‡	23.6	±0.3
24	Indiana		23.5	±0.2
23	Maine		23.4	±0.4
23	North Carolina		23.4	±0.2
23	Rhode Island	‡	23.4	±0.5
20	Missouri		23.1	±0.2
20	Ohio		23.1	±0.1
18	Minnesota		23.0	±0.2
17	Kentucky		22.9	±0.2
16	Oregon		22.5	±0.3
15	Vermont		21.9	±0.5
15	Wisconsin		21.9	±0.2
13	Utah		21.6	±0.3
12	New Mexico		21.4	±0.4
11	Arkansas		21.3	±0.4
10	Oklahoma		21.1	±0.2
9	Idaho		19.7	±0.4
8	Kansas		18.9	±0.3
7	Iowa		18.8	±0.2
6	Alaska		18.4	±0.5
5	Montana		18.2	±0.5
4	Nebraska		18.1	±0.3
4	Wyoming		18.1	±0.8
2	North Dakota		16.9	±0.6
2	South Dakota		16.9	±0.5

Early works

- Some of the early classical works on ranking and selection of population means are by Bechhofer (1954), Gupta (1956).
- A Bayesian approach to ranking several binomial populations: Bland and Bratcher (1968), Govindarajulu and Harvey (1974), Goel and Rubin (1977).
- Laird and Louis (1989) proposed an empirical Bayes (EB) approach, Morris and Christiansen (1994) proposed an approximate hierarchical Bayes (HB) approach.
- Aitkin and Longford (1986) and Laird and Louis (1989) used EB approach to ranking.
- Berger and Deely (1988) used an HB approach to ranking.

Outline

1. Introduction
2. Review of KWW (2020) Framework
3. Proposed Bayesian Framework
4. Baseball Example – Efron and Morris (1975)
5. Commuting Times Example – Klein et al. (2020)
6. Median Incomes Example
7. Summary and Conclusions

Estimation of overall ranking

- Klein et al. (2020) considered overall ranking of the 50 US states and DC, based on θ_i, the mean commuting times of the workers not working from home, $i = 1, \cdots, m = 51$.
- True means θ_i's are estimated from the ACS data; $y_1, \cdots, y_m$ are the estimates from ACS
- $\check{r}_1, \cdots, \check{r}_m$ are the ranks of the true means
- $R_1, \cdots, R_m$ are the ranks of $y_1, \cdots, y_m$
- Estimates of the ranks $\check{r}_1, \cdots, \check{r}_m$ from $R_1, \cdots, R_m$, the ranks of the point estimates $y_1, \cdots, y_m$ ignore the sampling error in the y-estimates

Klein et al. joint confidence set for θ

- KWW assumed the model

$$Y_i|\theta \overset{ind}{\sim} N(\theta_i, D_i),\ i = 1, \cdots, m,$$

- Based on this, KWW created $(1-\alpha)$-level *joint confidence set* for θ based on confidence intervals

$$I_i = \left(Y_i - z_{1-\frac{\gamma}{2}}\sqrt{D_i},\ Y_i - z_{\frac{\gamma}{2}}\sqrt{D_i}\right) \equiv (L_i, U_i),$$

  for $i = 1, \cdots, m$, for individual θ_i's. KWW determined γ from α by Bonferroni inequality or independence method.

- A joint $(1-\alpha)$-confidence set for θ is the Cartesian product of the intervals I_i, for $i = 1, \cdots, m$.

Klein et al. confidence solution for overall ranking

- From the joint confidence set $I_1 \times \cdots \times I_m$ of θ, KWW created confidence solution for overall ranking. For each $i \in \{1, 2, \cdots, m\}$, they define the sets

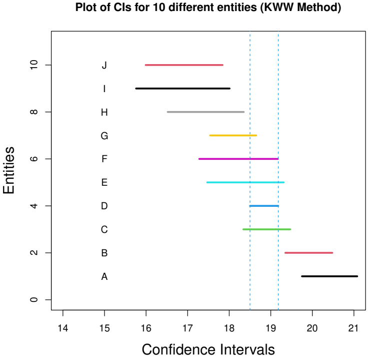
$$\begin{aligned} C_i &= \{1, 2, \cdots, m\} \setminus \{i\}, \\ \Lambda_{Li} &= \{j \in C_i : U_j \leq L_i\}, \\ \Lambda_{Ri} &= \{j \in C_i : U_i \leq L_j\}, \\ \Lambda_{Oi} &= \{j \in C_i : U_j > L_i \text{ and } U_i > L_j\} = C_i \setminus \{\Lambda_{Li} \cup \Lambda_{Ri}\} \end{aligned}$$

- Klein et al. (2020) show that the set of rank vectors

$$\left\{ (r_1, \cdots, r_m) : r_i \in \{|\Lambda_{Li}| + 1, \cdots, |\Lambda_{Li}| + 1 + |\Lambda_{Oi}|\} \text{ for } i = 1, \cdots, m \right\},$$

is a joint confidence set for the overall rank vector $\check{r} = (\check{r}_1, \cdots, \check{r}_m)$ with coverage probability at least $(1 - \alpha)$.

Illustration of KWW Ranking



Consider D $(i = 4)$. Then

$$\Lambda_{Li} = \{8, 9, 10\},\ \Lambda_{Ri} = \{1, 2\},\ \Lambda_{Oi} = \{3, 5, 6, 7\}$$

$$|\Lambda_{Li}| = 3,\ |\Lambda_{Oi}| = 4,\ r_i \in \{3+1, \cdots, 3+4+1\} = \{4, 5, 6, 7, 8\}$$

Joint conf. set for ranking from jt CIs: Table 2 of KWW

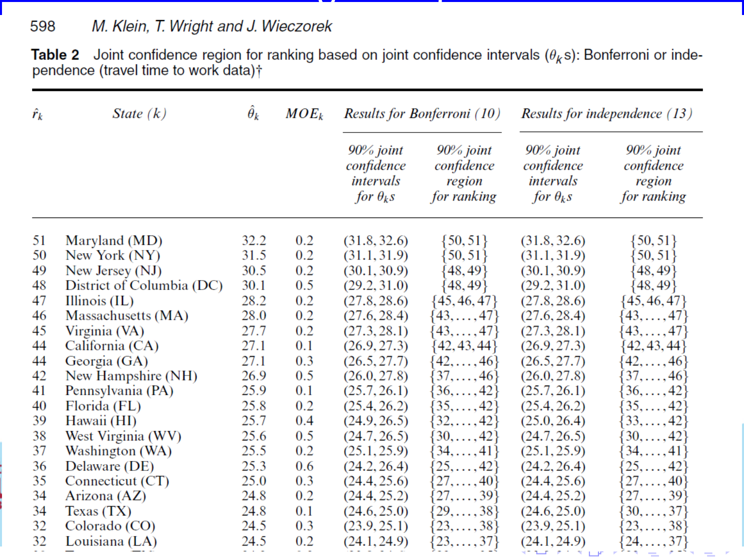
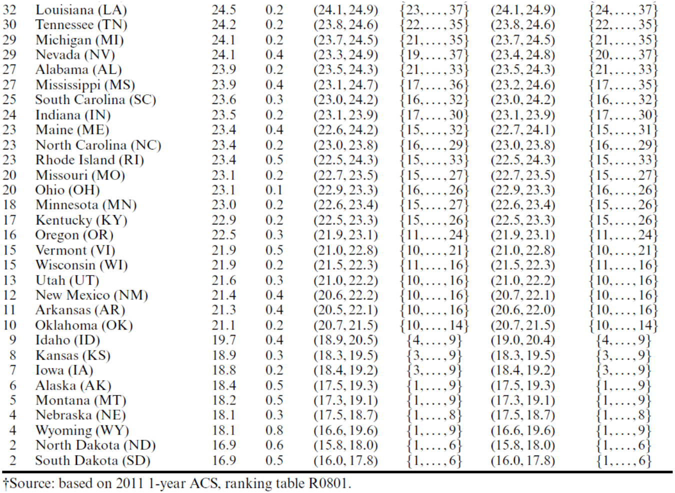


Table 2 Joint confidence region for ranking based on joint confidence intervals (θ_ks): Bonferroni or independence (travel time to work data)†

$\hat{r}_k$	*State (k)*	$\hat{\theta}_k$	MOE_k	*Results for Bonferroni (10)*		*Results for independence (13)*	
				90% joint confidence intervals for θ_ks	*90% joint confidence region for ranking*	*90% joint confidence intervals for θ_ks*	*90% joint confidence region for ranking*
51	Maryland (MD)	32.2	0.2	(31.8, 32.6)	{50, 51}	(31.8, 32.6)	{50, 51}
50	New York (NY)	31.5	0.2	(31.1, 31.9)	{50, 51}	(31.1, 31.9)	{50, 51}
49	New Jersey (NJ)	30.5	0.2	(30.1, 30.9)	{48, 49}	(30.1, 30.9)	{48, 49}
48	District of Columbia (DC)	30.1	0.5	(29.2, 31.0)	{48, 49}	(29.2, 31.0)	{48, 49}
47	Illinois (IL)	28.2	0.2	(27.8, 28.6)	{45, 46, 47}	(27.8, 28.6)	{45, 46, 47}
46	Massachusetts (MA)	28.0	0.2	(27.6, 28.4)	{43, ..., 47}	(27.6, 28.4)	{43, ..., 47}
45	Virginia (VA)	27.7	0.2	(27.3, 28.1)	{43, ..., 47}	(27.3, 28.1)	{43, ..., 47}
44	California (CA)	27.1	0.1	(26.9, 27.3)	{42, 43, 44}	(26.9, 27.3)	{42, 43, 44}
44	Georgia (GA)	27.1	0.3	(26.5, 27.7)	{42, ..., 46}	(26.5, 27.7)	{42, ..., 46}
42	New Hampshire (NH)	26.9	0.5	(26.0, 27.8)	{37, ..., 46}	(26.0, 27.8)	{37, ..., 46}
41	Pennsylvania (PA)	25.9	0.1	(25.7, 26.1)	{36, ..., 42}	(25.7, 26.1)	{36, ..., 42}
40	Florida (FL)	25.8	0.2	(25.4, 26.2)	{35, ..., 42}	(25.4, 26.2)	{35, ..., 42}
39	Hawaii (HI)	25.7	0.4	(24.9, 26.5)	{32, ..., 42}	(25.0, 26.4)	{33, ..., 42}
38	West Virginia (WV)	25.6	0.5	(24.7, 26.5)	{30, ..., 42}	(24.7, 26.5)	{30, ..., 42}
37	Washington (WA)	25.5	0.2	(25.1, 25.9)	{34, ..., 41}	(25.1, 25.9)	{34, ..., 41}
36	Delaware (DE)	25.3	0.6	(24.2, 26.4)	{25, ..., 42}	(24.2, 26.4)	{25, ..., 42}
35	Connecticut (CT)	25.0	0.3	(24.4, 25.6)	{27, ..., 40}	(24.4, 25.6)	{27, ..., 40}
34	Arizona (AZ)	24.8	0.2	(24.4, 25.2)	{27, ..., 39}	(24.4, 25.2)	{27, ..., 39}
34	Texas (TX)	24.8	0.1	(24.6, 25.0)	{29, ..., 38}	(24.6, 25.0)	{30, ..., 37}
32	Colorado (CO)	24.5	0.3	(23.9, 25.1)	{23, ..., 38}	(23.9, 25.1)	{23, ..., 38}
32	Louisiana (LA)	24.5	0.2	(24.1, 24.9)	{23, ..., 37}	(24.1, 24.9)	{24, ..., 37}

Joint conf. set for ranking from jt CIs: Table 2 of KWW



32	Louisiana (LA)	24.5	0.2	(24.1, 24.9)	{23, ..., 37}	(24.1, 24.9)	{24, ..., 37}
30	Tennessee (TN)	24.2	0.2	(23.8, 24.6)	{22, ..., 35}	(23.8, 24.6)	{22, ..., 35}
29	Michigan (MI)	24.1	0.2	(23.7, 24.5)	{21, ..., 35}	(23.7, 24.5)	{21, ..., 35}
29	Nevada (NV)	24.1	0.4	(23.3, 24.9)	{19, ..., 37}	(23.4, 24.8)	{20, ..., 37}
27	Alabama (AL)	23.9	0.2	(23.5, 24.3)	{21, ..., 33}	(23.5, 24.3)	{21, ..., 33}
27	Mississippi (MS)	23.9	0.4	(23.1, 24.7)	{17, ..., 36}	(23.2, 24.6)	{17, ..., 35}
25	South Carolina (SC)	23.6	0.3	(23.0, 24.2)	{16, ..., 32}	(23.0, 24.2)	{16, ..., 32}
24	Indiana (IN)	23.5	0.2	(23.1, 23.9)	{17, ..., 30}	(23.1, 23.9)	{17, ..., 30}
23	Maine (ME)	23.4	0.4	(22.6, 24.2)	{15, ..., 32}	(22.7, 24.1)	{15, ..., 31}
23	North Carolina (NC)	23.4	0.2	(23.0, 23.8)	{16, ..., 29}	(23.0, 23.8)	{16, ..., 29}
23	Rhode Island (RI)	23.4	0.5	(22.5, 24.3)	{15, ..., 33}	(22.5, 24.3)	{15, ..., 33}
20	Missouri (MO)	23.1	0.2	(22.7, 23.5)	{15, ..., 27}	(22.7, 23.5)	{15, ..., 27}
20	Ohio (OH)	23.1	0.1	(22.9, 23.3)	{16, ..., 26}	(22.9, 23.3)	{16, ..., 26}
18	Minnesota (MN)	23.0	0.2	(22.6, 23.4)	{15, ..., 27}	(22.6, 23.4)	{15, ..., 26}
17	Kentucky (KY)	22.9	0.2	(22.5, 23.3)	{15, ..., 26}	(22.5, 23.3)	{15, ..., 26}
16	Oregon (OR)	22.5	0.3	(21.9, 23.1)	{11, ..., 24}	(21.9, 23.1)	{11, ..., 24}
15	Vermont (VI)	21.9	0.5	(21.0, 22.8)	{10, ..., 21}	(21.0, 22.8)	{10, ..., 21}
15	Wisconsin (WI)	21.9	0.2	(21.5, 22.3)	{11, ..., 16}	(21.5, 22.3)	{11, ..., 16}
13	Utah (UT)	21.6	0.3	(21.0, 22.2)	{10, ..., 16}	(21.0, 22.2)	{10, ..., 16}
12	New Mexico (NM)	21.4	0.4	(20.6, 22.2)	{10, ..., 16}	(20.7, 22.1)	{10, ..., 16}
11	Arkansas (AR)	21.3	0.4	(20.5, 22.1)	{10, ..., 16}	(20.6, 22.0)	{10, ..., 16}
10	Oklahoma (OK)	21.1	0.2	(20.7, 21.5)	{10, ..., 14}	(20.7, 21.5)	{10, ..., 14}
9	Idaho (ID)	19.7	0.4	(18.9, 20.5)	{4, ..., 9}	(19.0, 20.4)	{4, ..., 9}
8	Kansas (KS)	18.9	0.3	(18.3, 19.5)	{3, ..., 9}	(18.3, 19.5)	{3, ..., 9}
7	Iowa (IA)	18.8	0.2	(18.4, 19.2)	{3, ..., 9}	(18.4, 19.2)	{3, ..., 9}
6	Alaska (AK)	18.4	0.5	(17.5, 19.3)	{1, ..., 9}	(17.5, 19.3)	{1, ..., 9}
5	Montana (MT)	18.2	0.5	(17.3, 19.1)	{1, ..., 9}	(17.3, 19.1)	{1, ..., 9}
4	Nebraska (NE)	18.1	0.3	(17.5, 18.7)	{1, ..., 8}	(17.5, 18.7)	{1, ..., 8}
4	Wyoming (WY)	18.1	0.8	(16.6, 19.6)	{1, ..., 9}	(16.6, 19.6)	{1, ..., 9}
2	North Dakota (ND)	16.9	0.6	(15.8, 18.0)	{1, ..., 6}	(15.8, 18.0)	{1, ..., 6}
2	South Dakota (SD)	16.9	0.5	(16.0, 17.8)	{1, ..., 6}	(16.0, 17.8)	{1, ..., 6}

†Source: based on 2011 1-year ACS, ranking table R0801.

Outline

1 Introduction

2 Review of KWW (2020) Framework

3 Proposed Bayesian Framework

4 Baseball Example – Efron and Morris (1975)

5 Commuting Times Example – Klein et al. (2020)

6 Median Incomes Example

7 Summary and Conclusions

Notion of a credible distribution of the overall ranking

- We construct Bayesian credible distributions of overall ranking by a sampling-based approach by drawing samples from the posterior distribution of θ.
- We generate a large sample $\mathcal{F}_M$, of size S, of $\theta_1, \cdots, \theta_m$ values from their joint posterior pdf, $\pi_M(\theta|y)$, derived under a model M.
- We empirically choose a suitable subsample of size $\approx S \times (1-\alpha)$ from these S samples; denote this set by $\mathcal{F}_{M,\alpha,y}$. This subsample corresponds to an appropriate credible set of θ.
- The set $\mathcal{F}_{M,\alpha,y}$ has empirical posterior probability $(1-\alpha)$. We rank each of the chosen θ and create a credible distribution for the true rank vector $\check{r}_1, \cdots, \check{r}_m$

Two Bayesian models

An unstructured Bayesian model:

(I) $Y_i|\theta_1,\cdots,\theta_m \overset{ind}{\sim} N(\theta_i,D_i), i=1,\cdots,m,$

(II) $\pi(\theta_1,\cdots,\theta_m)=1$ for $-\infty<\theta_1,\cdots,\theta_m<\infty$. The joint posterior pdf of θ is a known MVN.

Fay-Herriot model: a class of hierarchical Bayesian models:

(I) $Y_i|\theta_1,\cdots,\theta_m \overset{ind}{\sim} N(\theta_i,D_i), i=1,\cdots,m,$

(II) Conditional on model parameters β,A, subpopulation means θ_i's are independently distributed, given by $\theta_i|\beta,A \overset{ind}{\sim} N(x_i^T\beta,A), i=1,\cdots,m,$

(III) $\pi(\beta,A)=1/\sqrt{\bar{D}+A}$. Posterior distribution of θ can be easily sampled by MCMC or by non-Markovian method.

Credible set for θ: A Cartesian credible set

- From the sample in $\mathcal{F}_M$, we take the ith component of each vector to create a sample $\mathcal{F}_i$ for θ_i. For a suitable κ, we determine the $(\kappa/2)$th and $(1-\kappa/2)$th quantiles, a_i and b_i, of the set $\mathcal{F}_i$. Define now

$$\begin{aligned} \mathcal{S}_i &= \{s : a_i \leq \theta_i^{(s)} \leq b_i, s = 1, 2, \cdots S\} \\ \mathcal{S}_J &= \cap_{i=1}^{m} \mathcal{S}_i \end{aligned}$$

  Let $K_J = |\mathcal{S}_J|$. For any $\alpha \in (0,1)$, we determine κ in such a way that K_J is the integer closest to $S \times (1-\alpha)$.

- The selected set

$$\{\theta^{(s)}, s \in \mathcal{S}_J\}$$

  is an approximate representation of a $(1-\alpha)$ joint credible set for θ. We use these θ values to determine a rank distribution, described below.

A nearly optimal credible set for θ: an elliptical set

- An optimal method to create credible set for θ is the HPD method.
  - For the UB model, HPD credible set is elliptical.
  - For the HB model, an HPD set is approximately elliptical.
- The HPD credible set is centered at

$$\hat{\theta} = \frac{1}{S}\sum_{s=1}^{S} \theta^{(s)}$$

and the elliptical shape is determined by the dispersion matrix

$$V = \frac{1}{S}\sum_{s=1}^{S} \{\theta^{(s)} - \hat{\theta}\}\{\theta^{(s)} - \hat{\theta}\}^T.$$

A nearly optimal credible set for θ: an elliptical set

- Create Mahalanobis distances
$$d_B^{(s)} = M\left(\theta^{(s)}, \hat{\theta}, V\right), \text{ for } s = 1, \cdots, S.$$
Find c, the $(1-\alpha)$-quantile of these distances, and create an empirical credible set $\mathcal{F}_{M,\alpha,y}$ for θ with $\{s \in \mathcal{F}_M : d_B^{(s)} \leq c\}$.
- The selected set $\mathcal{F}_{M,\alpha,y}$ is an approximate representation of a $(1-\alpha)$ joint HPD credible set for θ.
- Again, we use these θ values to determine a rank distribution below.

Joint credible distribution of overall ranking

- For each sample in $\mathcal{F}_{M,\alpha,y}$, we create a two-way $m \times m$ table, columns as the populations or subjects to be ranked and rows as the ranks of the components of θ, explained now. For each sample such as $\theta^{(s_1)}$, we start with an $m \times m$ null matrix.
- If the ith component of $\theta^{(s_1)}$ does not tie with any other component and has rank j_i, we put a 1 in the j_ith row of the ith column. If two components k and l tie for ranks j and $j+1$, we replace elements $(k,j),(k,j+1),(l,j),(l,j+1)$ by $1/2$.
- For all K elements in $\mathcal{F}_{M,\alpha,y}$, we complete K tables, and for some weight $w(\theta)$ we take a weighted average of these tables, where the weights sum to 1. This produces a credible distribution of overall ranking. We denote the distribution of $\check{r}_i$ by $\pi_{\check{r}_i,M,T,y}$.

Outline

1 Introduction

2 Review of KWW (2020) Framework

3 Proposed Bayesian Framework

4 Baseball Example – Efron and Morris (1975)

5 Commuting Times Example – Klein et al. (2020)

6 Median Incomes Example

7 Summary and Conclusions

Ranking baseball players: An example by Efron and Morris

- Efron and Morris (1975) considered batting averages of 18 major league baseball batters from their first 45 at bats in the 1970 baseball season to predict their performances in the remainder of that season.
- A reasonable representative value for θ_i was the player's average (ω_i), known, in the remainder of the season after his first 45 at bats.
- The sample proportion of hits Y_i from first 45 at bats is approximately normal with mean θ_i and estimated variance $D_i = Y_i(1 - Y_i)/45$.

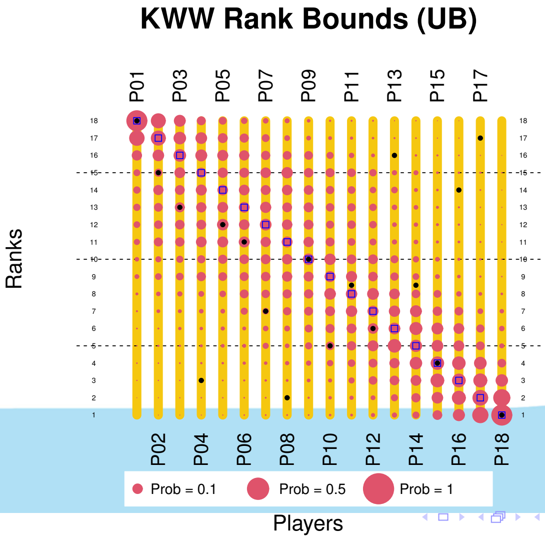
KWW confidence set of overall ranking of players

- From equation (9) of Klein et al. (2020) a 90% joint confidence region for the rank vector $\check{r}$ is given by

$$\{(\check{r}_1, \cdots, \check{r}_{18}) : \check{r}_k \in \{1, 2, \cdots, 18\} \text{ for } k = 1, \cdots, 18\}.$$

  - This confidence set for the rank vector is depicted in the Figure next slide.
  - Possible ranks for each player in this set are shown by a yellow line segment stretching from 1 to 18.
  - According to this confidence set, any player can rank from 1 to 18, and any particular ranks can be held by any of these 18 players.
- The credible distribution is pictorially presented by overlaying on the figure of the confidence set solid red circles with area of a circle is proportional to the probability it is representing.

Conf sets and credible distrns (Baseball)



Comparison of confidence sets and credible distributions

- ω_i: Batting average from the remainder of the season after first 45 at bats for player i.
- ξ_i: A surrogate of true rank based on ω_i.
- Compute for player $i = 1, \cdots, 18$,

$$\varepsilon_{i,HB,C} = E^{\pi_{\check{r}_i,HB,C,y}}\left(|\check{r}_i - \xi_i| \middle| y\right), \qquad \varepsilon_{i,HB,E} = E^{\pi_{\check{r}_i,HB,E,y}}\left(|\check{r}_i - \xi_i| \middle| y\right),$$

$$\varepsilon_{i,UB,C} = E^{\pi_{\check{r}_i,UB,C,y}}\left(|\check{r}_i - \xi_i| \middle| y\right), \qquad \varepsilon_{i,UB,E} = E^{\pi_{\check{r}_i,UB,E,y}}\left(|\check{r}_i - \xi_i| \middle| y\right),$$

$$\varepsilon_{i,KWW} = \frac{1}{|\Lambda_{Oi}|+1} \sum_{j=|\Lambda_{Li}|+1}^{|\Lambda_{Li}|+|\Lambda_{Oi}|+1} |j - \xi_i|.$$

Application to 1970 Batting averages of 18 Major League players

Players	y_i	ω_i	R_i	ξ_i	$\varepsilon_{i,HB,C}$	$\varepsilon_{i,HB,E}$	$\varepsilon_{i,UB,C}$	$\varepsilon_{i,UB,E}$	$\varepsilon_{i,KWW}$
1. Clemente	.400	.346	18	18	5.90	5.90	1.18	1.18	8.50
2. F. Robinson	.378	.298	17	15	4.39	4.39	1.87	1.87	6.17
3. F. Howard	.356	.276	16	13	4.24	21.0	2.73	2.73	5.17
4. Johnstone	.333	.222	15	3	8.11	8.11	11.40	11.39	**6.83**
5. Berry	.311	.273	13.5	12	4.33	4.33	2.44	2.44	4.83
6. Spencer	.311	.270	13.5	11	4.25	4.25	2.96	2.96	4.61
7. Kessinger	.289	.263	12	7	4.96	4.96	4.92	4.92	4.83
8. L. Alvarado	.267	.210	11	2	7.66	7.66	8.13	8.13	**7.61**
9. Santo	.244	.269	9.5	10	4.36	4.34	3.09	3.09	4.50
10. Swoboda	.244	.230	9.5	5	5.41	5.41	3.98	3.98	5.61
11. Unser	.222	.264	6	8.5	4.26	4.26	3.09	3.08	4.56
12. Williams	.222	.256	6	6	4.40	4.40	2.64	2.64	5.17
13. Scott	.222	.303	6	16	7.64	7.64	9.36	9.36	**6.83**
14. Petrocelli	.222	.264	6	8.5	4.25	4.25	3.33	3.33	4.56
15. E. Rodriguez	.222	.226	6	4	5.38	5.38	3.46	3.46	6.17
16. Campenaris	.200	.285	3	14	6.64	6.64	8.97	8.97	**5.61**
17. Munson	.178	.316	2	17	9.54	9.54	13.57	13.56	**7.61**
18. Alvis	.156	.200	1	1	5.46	5.46	1.43	1.43	8.50
Total abs. dev.					101.16	101.16	88.56	88.56	107.64
$10^{13}\times$ Vol.					51.38	1.32	93756	1757	85722
Average length					0.236	0.193	0.359	0.288	0.357

Summary of simulations for the baseball setup

1:=APEAD(W), 2:=APEAD(UW), 3:=$(AveVol)^{1/m}$, 4:=Ave length

		With covariate					No covariate				
			H. Bayes		U. Bayes			H. Bayes		U. Bayes	
A		KWW	Cart	Ellip	Cart	Ellip	KWW	Cart	Ellip	Cart	Ellip
0.001	1	5.518	1.178	1.178	2.313	2.313	5.978	5.366	5.366	4.642	4.642
	2	5.518	1.543	1.521	2.560	2.545	5.978	5.194	5.199	4.804	4.795
	3	0.356	0.262	0.212	0.357	0.286	0.356	0.259	0.209	0.357	0.287
	4	0.357	0.247	0.199	0.358	0.287	0.357	0.241	0.196	0.358	0.287
0.005	1	5.268	1.835	1.835	2.148	2.148	5.919	4.065	4.066	3.468	3.468
	2	5.268	2.033	2.020	2.378	2.365	5.919	4.094	4.094	3.714	3.701
	3	0.356	0.297	0.238	0.357	0.287	0.356	0.295	0.237	0.356	0.286
	4	0.357	0.289	0.232	0.358	0.287	0.357	0.281	0.227	0.358	0.287
0.01	1	4.879	1.874	1.874	1.980	1.980	5.668	2.967	2.967	2.766	2.766
	2	4.879	2.063	2.053	2.188	2.176	5.668	3.211	3.203	3.013	2.999
	3	0.356	0.318	0.255	0.357	0.286	0.356	0.317	0.254	0.357	0.287
	4	0.357	0.314	0.252	0.358	0.287	0.357	0.312	0.251	0.358	0.288
0.1	1	2.919	1.070	1.070	1.061	1.061	3.212	1.177	1.177	1.168	1.168
	2	2.919	1.205	1.198	1.194	1.186	3.212	1.328	1.319	1.313	1.305
	3	0.356	0.351	0.282	0.356	0.287	0.356	0.350	0.281	0.356	0.286
	4	0.357	0.352	0.283	0.358	0.287	0.357	0.351	0.282	0.358	0.287
1	1	1.209	0.417	0.417	0.418	0.418	1.181	0.370	0.370	0.369	0.369
	2	1.209	0.467	0.465	0.467	0.464	1.181	0.426	0.423	0.426	0.423
	3	0.356	0.356	0.286	0.357	0.287	0.356	0.356	0.286	0.357	0.287
	4	0.357	0.357	0.287	0.358	0.288	0.357	0.357	0.287	0.358	0.288

Outline

1 Introduction

2 Review of KWW (2020) Framework

3 Proposed Bayesian Framework

4 Baseball Example – Efron and Morris (1975)

5 Commuting Times Example – Klein et al. (2020)

6 Median Incomes Example

7 Summary and Conclusions

Ranking of US states based on commuting times

- Recall that, in a pioneering article, Klein et al. (2020) applied their frequentist approach to rank fifty states of the U.S. and DC by mean commuting times of workers sixteen or older and not working from home. They used survey data collected from the ACS.
- Let us revisit their Table 2 next.

Joint conf. set for ranking from jt CIs: Table 2 of KWW

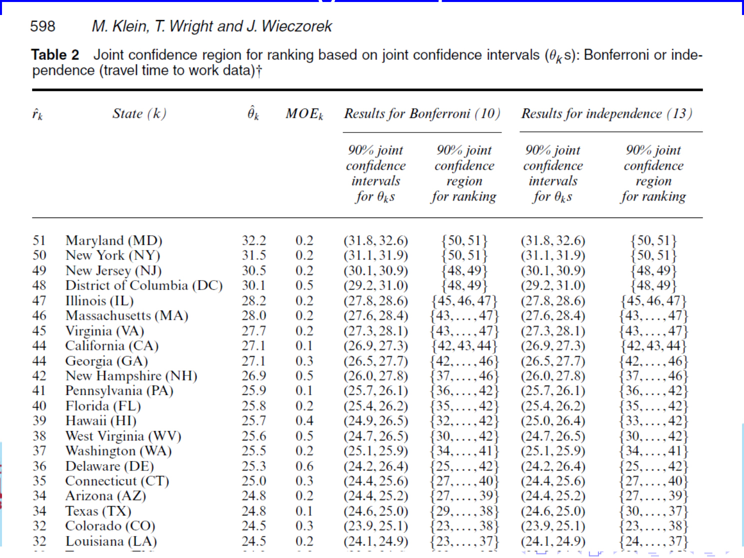
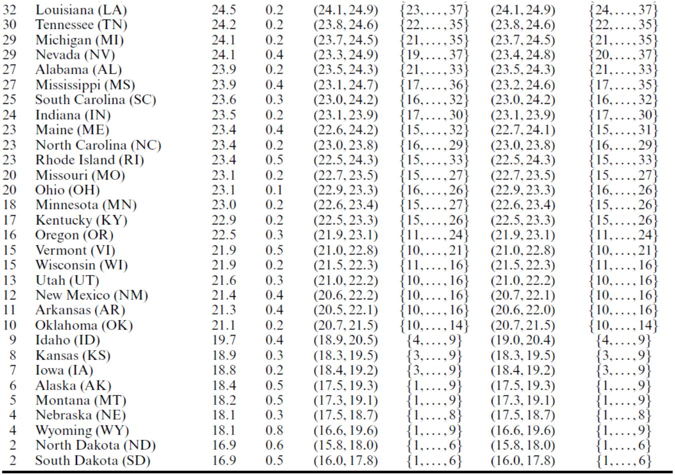


Table 2 Joint confidence region for ranking based on joint confidence intervals (θ_ks): Bonferroni or independence (travel time to work data)†

$\hat{r}_k$	*State (k)*	$\hat{\theta}_k$	MOE_k	*Results for Bonferroni (10)*		*Results for independence (13)*	
				90% joint confidence intervals for θ_ks	*90% joint confidence region for ranking*	*90% joint confidence intervals for θ_ks*	*90% joint confidence region for ranking*
51	Maryland (MD)	32.2	0.2	(31.8, 32.6)	{50, 51}	(31.8, 32.6)	{50, 51}
50	New York (NY)	31.5	0.2	(31.1, 31.9)	{50, 51}	(31.1, 31.9)	{50, 51}
49	New Jersey (NJ)	30.5	0.2	(30.1, 30.9)	{48, 49}	(30.1, 30.9)	{48, 49}
48	District of Columbia (DC)	30.1	0.5	(29.2, 31.0)	{48, 49}	(29.2, 31.0)	{48, 49}
47	Illinois (IL)	28.2	0.2	(27.8, 28.6)	{45, 46, 47}	(27.8, 28.6)	{45, 46, 47}
46	Massachusetts (MA)	28.0	0.2	(27.6, 28.4)	{43, . . . , 47}	(27.6, 28.4)	{43, . . . , 47}
45	Virginia (VA)	27.7	0.2	(27.3, 28.1)	{43, . . . , 47}	(27.3, 28.1)	{43, . . . , 47}
44	California (CA)	27.1	0.1	(26.9, 27.3)	{42, 43, 44}	(26.9, 27.3)	{42, 43, 44}
44	Georgia (GA)	27.1	0.3	(26.5, 27.7)	{42, . . . , 46}	(26.5, 27.7)	{42, . . . , 46}
42	New Hampshire (NH)	26.9	0.5	(26.0, 27.8)	{37, . . . , 46}	(26.0, 27.8)	{37, . . . , 46}
41	Pennsylvania (PA)	25.9	0.1	(25.7, 26.1)	{36, . . . , 42}	(25.7, 26.1)	{36, . . . , 42}
40	Florida (FL)	25.8	0.2	(25.4, 26.2)	{35, . . . , 42}	(25.4, 26.2)	{35, . . . , 42}
39	Hawaii (HI)	25.7	0.4	(24.9, 26.5)	{32, . . . , 42}	(25.0, 26.4)	{33, . . . , 42}
38	West Virginia (WV)	25.6	0.5	(24.7, 26.5)	{30, . . . , 42}	(24.7, 26.5)	{30, . . . , 42}
37	Washington (WA)	25.5	0.2	(25.1, 25.9)	{34, . . . , 41}	(25.1, 25.9)	{34, . . . , 41}
36	Delaware (DE)	25.3	0.6	(24.2, 26.4)	{25, . . . , 42}	(24.2, 26.4)	{25, . . . , 42}
35	Connecticut (CT)	25.0	0.3	(24.4, 25.6)	{27, . . . , 40}	(24.4, 25.6)	{27, . . . , 40}
34	Arizona (AZ)	24.8	0.2	(24.4, 25.2)	{27, . . . , 39}	(24.4, 25.2)	{27, . . . , 39}
34	Texas (TX)	24.8	0.1	(24.6, 25.0)	{29, . . . , 38}	(24.6, 25.0)	{30, . . . , 37}
32	Colorado (CO)	24.5	0.3	(23.9, 25.1)	{23, . . . , 38}	(23.9, 25.1)	{23, . . . , 38}
32	Louisiana (LA)	24.5	0.2	(24.1, 24.9)	{23, . . . , 37}	(24.1, 24.9)	{24, . . . , 37}

Joint conf. set for ranking from jt CIs: Table 2 of KWW



32	Louisiana (LA)	24.5	0.2	(24.1, 24.9)	{23, . . . , 37}	(24.1, 24.9)	{24, . . . , 37}
30	Tennessee (TN)	24.2	0.2	(23.8, 24.6)	{22, . . . , 35}	(23.8, 24.6)	{22, . . . , 35}
29	Michigan (MI)	24.1	0.2	(23.7, 24.5)	{21, . . . , 35}	(23.7, 24.5)	{21, . . . , 35}
29	Nevada (NV)	24.1	0.4	(23.3, 24.9)	{19, . . . , 37}	(23.4, 24.8)	{20, . . . , 37}
27	Alabama (AL)	23.9	0.2	(23.5, 24.3)	{21, . . . , 33}	(23.5, 24.3)	{21, . . . , 33}
27	Mississippi (MS)	23.9	0.4	(23.1, 24.7)	{17, . . . , 36}	(23.2, 24.6)	{17, . . . , 35}
25	South Carolina (SC)	23.6	0.3	(23.0, 24.2)	{16, . . . , 32}	(23.0, 24.2)	{16, . . . , 32}
24	Indiana (IN)	23.5	0.2	(23.1, 23.9)	{17, . . . , 30}	(23.1, 23.9)	{17, . . . , 30}
23	Maine (ME)	23.4	0.4	(22.6, 24.2)	{15, . . . , 32}	(22.7, 24.1)	{15, . . . , 31}
23	North Carolina (NC)	23.4	0.2	(23.0, 23.8)	{16, . . . , 29}	(23.0, 23.8)	{16, . . . , 29}
23	Rhode Island (RI)	23.4	0.5	(22.5, 24.3)	{15, . . . , 33}	(22.5, 24.3)	{15, . . . , 33}
20	Missouri (MO)	23.1	0.2	(22.7, 23.5)	{15, . . . , 27}	(22.7, 23.5)	{15, . . . , 27}
20	Ohio (OH)	23.1	0.1	(22.9, 23.3)	{16, . . . , 26}	(22.9, 23.3)	{16, . . . , 26}
18	Minnesota (MN)	23.0	0.2	(22.6, 23.4)	{15, . . . , 27}	(22.6, 23.4)	{15, . . . , 26}
17	Kentucky (KY)	22.9	0.2	(22.5, 23.3)	{15, . . . , 26}	(22.5, 23.3)	{15, . . . , 26}
16	Oregon (OR)	22.5	0.3	(21.9, 23.1)	{11, . . . , 24}	(21.9, 23.1)	{11, . . . , 24}
15	Vermont (VI)	21.9	0.5	(21.0, 22.8)	{10, . . . , 21}	(21.0, 22.8)	{10, . . . , 21}
15	Wisconsin (WI)	21.9	0.2	(21.5, 22.3)	{11, . . . , 16}	(21.5, 22.3)	{11, . . . , 16}
13	Utah (UT)	21.6	0.3	(21.0, 22.2)	{10, . . . , 16}	(21.0, 22.2)	{10, . . . , 16}
12	New Mexico (NM)	21.4	0.4	(20.6, 22.2)	{10, . . . , 16}	(20.7, 22.1)	{10, . . . , 16}
11	Arkansas (AR)	21.3	0.4	(20.5, 22.1)	{10, . . . , 16}	(20.6, 22.0)	{10, . . . , 16}
10	Oklahoma (OK)	21.1	0.2	(20.7, 21.5)	{10, . . . , 14}	(20.7, 21.5)	{10, . . . , 14}
9	Idaho (ID)	19.7	0.4	(18.9, 20.5)	{4, . . . , 9}	(19.0, 20.4)	{4, . . . , 9}
8	Kansas (KS)	18.9	0.3	(18.3, 19.5)	{3, . . . , 9}	(18.3, 19.5)	{3, . . . , 9}
7	Iowa (IA)	18.8	0.2	(18.4, 19.2)	{3, . . . , 9}	(18.4, 19.2)	{3, . . . , 9}
6	Alaska (AK)	18.4	0.5	(17.5, 19.3)	{1, . . . , 9}	(17.5, 19.3)	{1, . . . , 9}
5	Montana (MT)	18.2	0.5	(17.3, 19.1)	{1, . . . , 9}	(17.3, 19.1)	{1, . . . , 9}
4	Nebraska (NE)	18.1	0.3	(17.5, 18.7)	{1, . . . , 8}	(17.5, 18.7)	{1, . . . , 8}
4	Wyoming (WY)	18.1	0.8	(16.6, 19.6)	{1, . . . , 9}	(16.6, 19.6)	{1, . . . , 9}
2	North Dakota (ND)	16.9	0.6	(15.8, 18.0)	{1, . . . , 6}	(15.8, 18.0)	{1, . . . , 6}
2	South Dakota (SD)	16.9	0.5	(16.0, 17.8)	{1, . . . , 6}	(16.0, 17.8)	{1, . . . , 6}

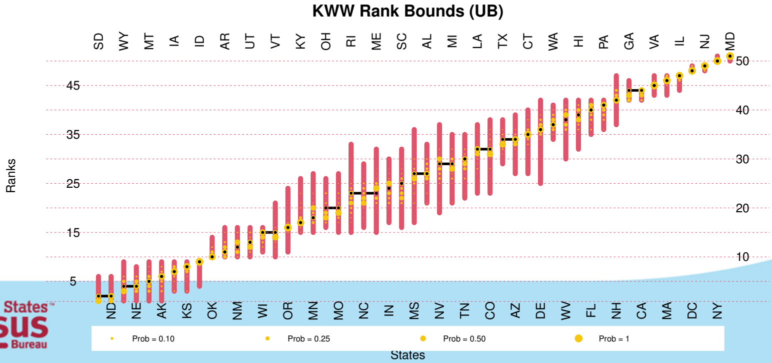
†Source: based on 2011 1-year ACS, ranking table R0801.

Ranking of US states based on commuting times

- In the figure below, we recreated Figure 1 of Klein et al. (2020) that depicted the frequentist solution of the confidence region for ranking.
- From the figure, for example, the possible ranks from this solution for the state ID are $4-9$.
- On the other hand, the states which can hold the rank 9 are the states WY, AK, MT, IA, KS and ID (the pink line segments for these states intersect the horizontal line for ranks = 9).

KWW2020 Figure 1

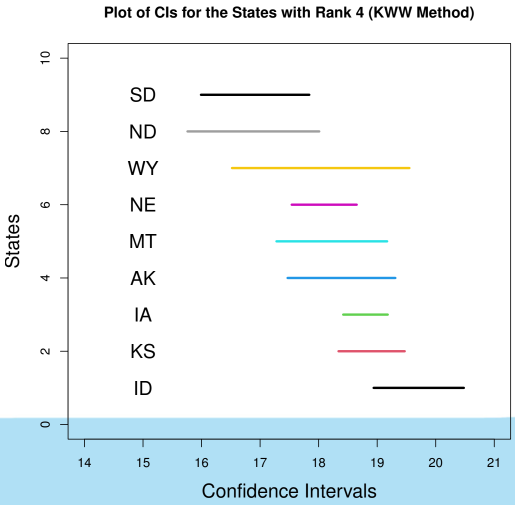
An alternative visualization of the 90% joint confidence region for travel time ranking given by Figure 1 of Klein et al. (2020) with an overlapping credible distribution from unstructured Bayes method



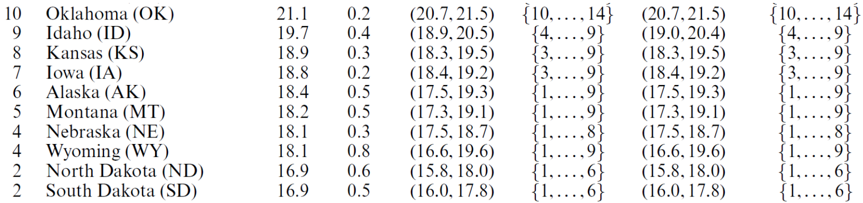
Credible distribution from Cartesian credible intervals

- On the last figure we overlaid credible distribution based on a C. credible set from the UB model. Probabilities from credible distribution are shown by yellow circles, bigger circles for larger probabilities.
- To interpret probabilities depicted in the figure, we focus specifically on its 4th row and 4th column. From the 4th column of this figure we find that the state of NE can have ranks 3$-$6 with respective probabilities about $0.2, 0.5, 0.2$, and 0.1.
- The 4th column in the figure shows the rank set for NE based on Klein et al. (2020) solution. It shows that in addition to the four ranks 3 to 6, NE can also rank 1, 2, 7 and 8.
- For ID, 9th column shows the KWW rank set $\{4, \cdots, 9\}$. Rank 9 on the 9th row is captured by one of six states: WY, MT, IA, AK, KS, ID. However, credible probabilities for ID and for rank 9 are 1.

CI’s for means of certain contender states (Rank = 4)



Nine states likely at rank 4 (Q4 of Table 2 from KWW)



10	Oklahoma (OK)	21.1	0.2	(20.7, 21.5)	$\{10, \ldots, 14\}$	(20.7, 21.5)	$\{10, \ldots, 14\}$
9	Idaho (ID)	19.7	0.4	(18.9, 20.5)	$\{4, \ldots, 9\}$	(19.0, 20.4)	$\{4, \ldots, 9\}$
8	Kansas (KS)	18.9	0.3	(18.3, 19.5)	$\{3, \ldots, 9\}$	(18.3, 19.5)	$\{3, \ldots, 9\}$
7	Iowa (IA)	18.8	0.2	(18.4, 19.2)	$\{3, \ldots, 9\}$	(18.4, 19.2)	$\{3, \ldots, 9\}$
6	Alaska (AK)	18.4	0.5	(17.5, 19.3)	$\{1, \ldots, 9\}$	(17.5, 19.3)	$\{1, \ldots, 9\}$
5	Montana (MT)	18.2	0.5	(17.3, 19.1)	$\{1, \ldots, 9\}$	(17.3, 19.1)	$\{1, \ldots, 9\}$
4	Nebraska (NE)	18.1	0.3	(17.5, 18.7)	$\{1, \ldots, 8\}$	(17.5, 18.7)	$\{1, \ldots, 8\}$
4	Wyoming (WY)	18.1	0.8	(16.6, 19.6)	$\{1, \ldots, 9\}$	(16.6, 19.6)	$\{1, \ldots, 9\}$
2	North Dakota (ND)	16.9	0.6	(15.8, 18.0)	$\{1, \ldots, 6\}$	(15.8, 18.0)	$\{1, \ldots, 6\}$
2	South Dakota (SD)	16.9	0.5	(16.0, 17.8)	$\{1, \ldots, 6\}$	(16.0, 17.8)	$\{1, \ldots, 6\}$

Credible Distrns of Ranks vs. Conf. Sets (a portion)

Commuting time example: Comparison of credible distributions of ranks for states implied by unstructured Cartesian sets with the joint confidence sets of ranks by KWW. States considered are those which are contenders of rank 4 by KWW solution.

Rank	**ID**	**KS**	**IA**	**AK**	**MT**	NE	**WY**	ND	SD
1							0.01	37.37	62.62
2							0.15	62.58	37.27
3				1.76	16.40	20.86	60.81	0.06	0.11
4				3.91	22.53	55.41	18.15		
5		0.79	0.37	13.81	51.67	20.10	13.26		
6		1.75	4.33	76.16	9.03	3.64	5.10		
7		13.10	81.30	3.91	0.34		1.35		
8		84.35	14.00	0.46	0.03		1.16		
9	100								
Conf. set	4-9	3-9	3-9	1-9	1-9	1-8	1-9	1-6	1-6
Mean	19.7	18.9	18.8	18.4	18.2	18.1	18.1	16.9	16.9
MOE	0.749	0.562	0.375	0.936	0.936	0.562	1.498	1.124	0.936

Outline

1 Introduction

2 Review of KWW (2020) Framework

3 Proposed Bayesian Framework

4 Baseball Example – Efron and Morris (1975)

5 Commuting Times Example – Klein et al. (2020)

6 Median Incomes Example

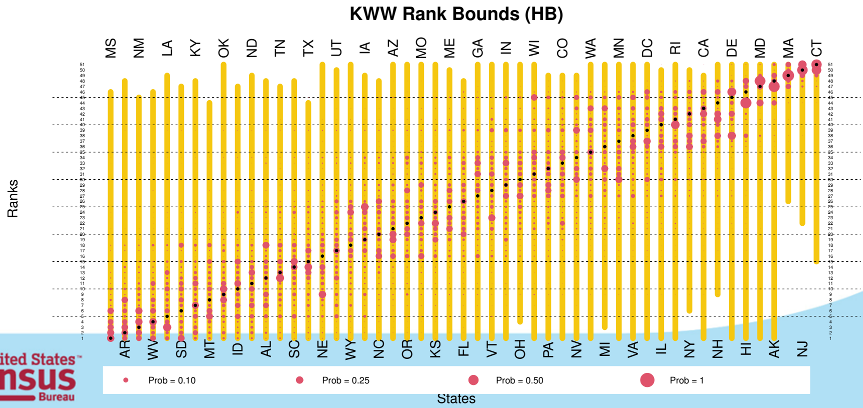
7 Summary and Conclusions

Ranking of US states based on median incomes

- Inference on median incomes of the US states for the income year 1989 has served as a benchmark example to evaluate effectiveness of various small area estimation methods.
- Fay (1987) suggested using two covariates for Fay-Herriot model. x_{i1} is the ith state median income for 1979 from 1980 Census, and $x_{i2} = (\mathrm{PCI}_{i,1989}/\mathrm{PCI}_{i,1979})x_{i1}$, $i = 1,\ldots,m$, where $\mathrm{PCI}_{i,1979}$ 1979 per capita income of the ith state.
- The 1990 census incomes for states were medians based on a large number of households from each state. These are “gold standard”. This will be used to create “surrogate ranks” ξ_i for true ranks $\check{r}$.

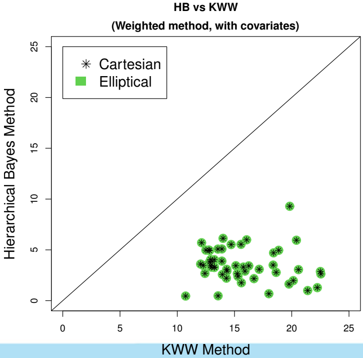
A visualization of median income ranking

90% joint conf region for median income ranking by KWW with an overlapping credible distrn by HB method

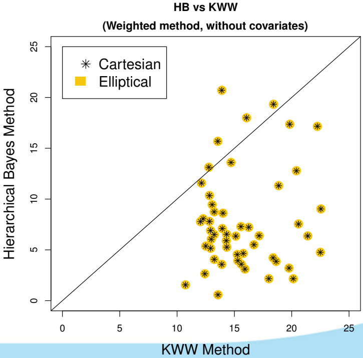


Plots of expected or average absolute deviations

HB vs. KWW



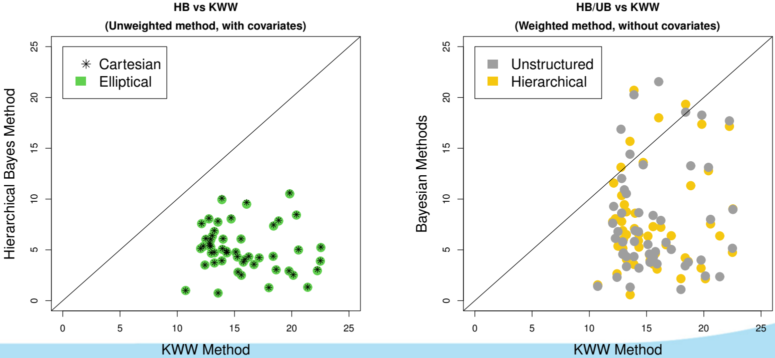
With Covariates



Covariates not used

Plots of expected or ave abs deviations

HB/UB vs. KWW

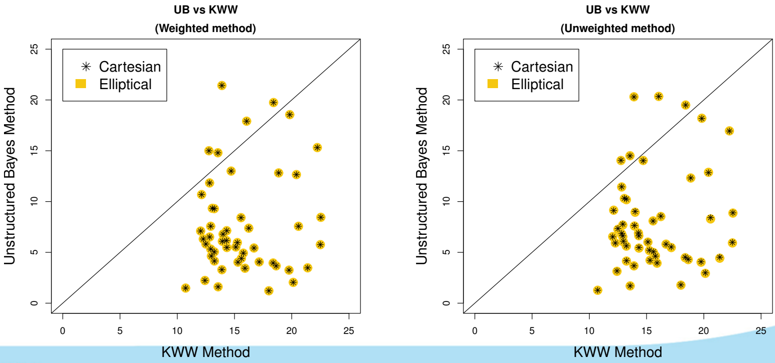


With Covariates

No covariates, elliptical

Plots of expected or average absolute deviations

UB-W vs. KWW and UB-UW vs. KWW



Covariates cannot be used Covariates cannot be used

Outline

1. Introduction
2. Review of KWW (2020) Framework
3. Proposed Bayesian Framework
4. Baseball Example – Efron and Morris (1975)
5. Commuting Times Example – Klein et al. (2020)
6. Median Incomes Example
7. Summary and Conclusions

Concluding Remarks

- In small area estimation, there is also some interest in estimating the ranks of the individual small areas based on their means. Shen and Louis (1998, JRSS B), for example, simultaneously considered estimation of small area means and point estimation of ranks.
- Accounting error of estimation of ranks is also important. To address this challenge, in a pioneering article Klein et al. (2020) considered set estimation of overall ranking. They created joint confidence set of the rank vector by employing a rectangular confidence set of the mean vector θ. The solution finds lower and upper bounds as the possible values for the true rank $\check{r}_i, i = 1, \cdots, m$. This frequentist solution may be interpreted as a confidence distribution for $\check{r}_i$, uniformly distributed over the integers between the lower and the upper bounds presented above.

Concluding Remarks

- The frequentist method cannot use auxiliary information or model the unknown true mean θ. Our Bayesian approach allows modeling of the mean vector to utilize useful auxiliary information as well as to develop the various credible sets for θ including the HPD sets, exact or nearly exact.
- We use appropriate credible sets to create a credible distribution for the true rank $\check{r}_i$. It is more informative than the frequentist confidence set. In the Bayesian proposal we can also allow more probable θ's in the credible set to have them bigger weights to the credible distribution. Application of proposed credible distributions is computationally straightforward.

Concluding Remarks

- Demonstrated our solution to three applications. All three applications and a simulation study mimicking two of the applications demonstrated superior performance of the proposed method over the existing frequentist method.
- Evaluation of frequentist confidence set and the credible distributions based on average or expected absolute deviations of estimated ranks from the “surrogate ranks” show better performance for the Bayesian solutions. With no covariates UB credible distributions perform better than HB credible distributions. Commuting example neither has any covariates nor any representative values of the unknown true means. Even then the credible distributions are “more conclusive”, “less dispersed” and “more illuminating” than the frequentist solution.

Concluding Remarks

- The median income problem has useful covariates as well as representative values (from the 1990 census) of the true means. Credible distributions from the HB model with the covariates performed the best when measured in terms average absolute deviations. Both credible distributions, from the UB model and the HB model, have better performance than the frequentist method.
- The HB credible distribution with covariates is uniformly better than the other solutions. But, the without covariates it does not have uniform superiority. Lack of covariates degrades the HB credible distribution.
- In terms of size measures of the confidence or credible sets for θ, Bayesian solutions are vastly superior, and HB credible sets are shorter or more “compact” than the UB-based credible sets.

In Summary,

- We proposed a Bayesian formulation for inference on overall ranking of a set of entities
- It is competitive with recent frequentist methods, and more effective and informative, and is as easy to implement as it is to compute the posterior means and variances of the entity means.
- Using credible sets, we created novel credible distributions for the rank vector of the entities.
- We evaluate the Bayesian procedure in terms of accuracy and stability in two applications and a simulation study.
- Frequentist approaches cannot take account of covariates, but the Bayesian method handles them easily

THANK YOU!!!

References:

1. Klein, M., Wright, T. and Wieczorek, J. (2020). “A Joint Confidence Region for an Overall Ranking of Populations”, *Journal of the Royal Statistical Society (Series C)*, 69, Part 3, 589–606.
2. Datta, G.; Hou, Y. and Mandal, A. (2024), “Credible Distributions of Overall Ranking of Entities”, submitted. `https://arxiv.org/pdf/2401.01833.pdf`

IAOS-ISI 2024
MEXICO CONFERENCE

Statistical Data Integration Using Multilevel Models to Predict Employee Compensation

Andreea Erciulescu
Jean Opsomer
Benjamin Schneider, Westat

INEGI

International Statistical Institute

IAOS
IMPROVING OFFICIAL STATISTICS

Motivation

- U.S. Bureau of Labor Statistics (BLS) sought to estimate total employee compensation (wages *and* benefits) for detailed occupations within small geographies
  - Example:
    - Construction laborers in the New York metropolitan statistical area have average wages of $26 / hour and benefits worth $5 / hour

- Example applications:
  - Pricing of labor for contracts in public works projects
  - Economic development research for cities and regions

Domains of Interest: Occupation by Geography

Occupations

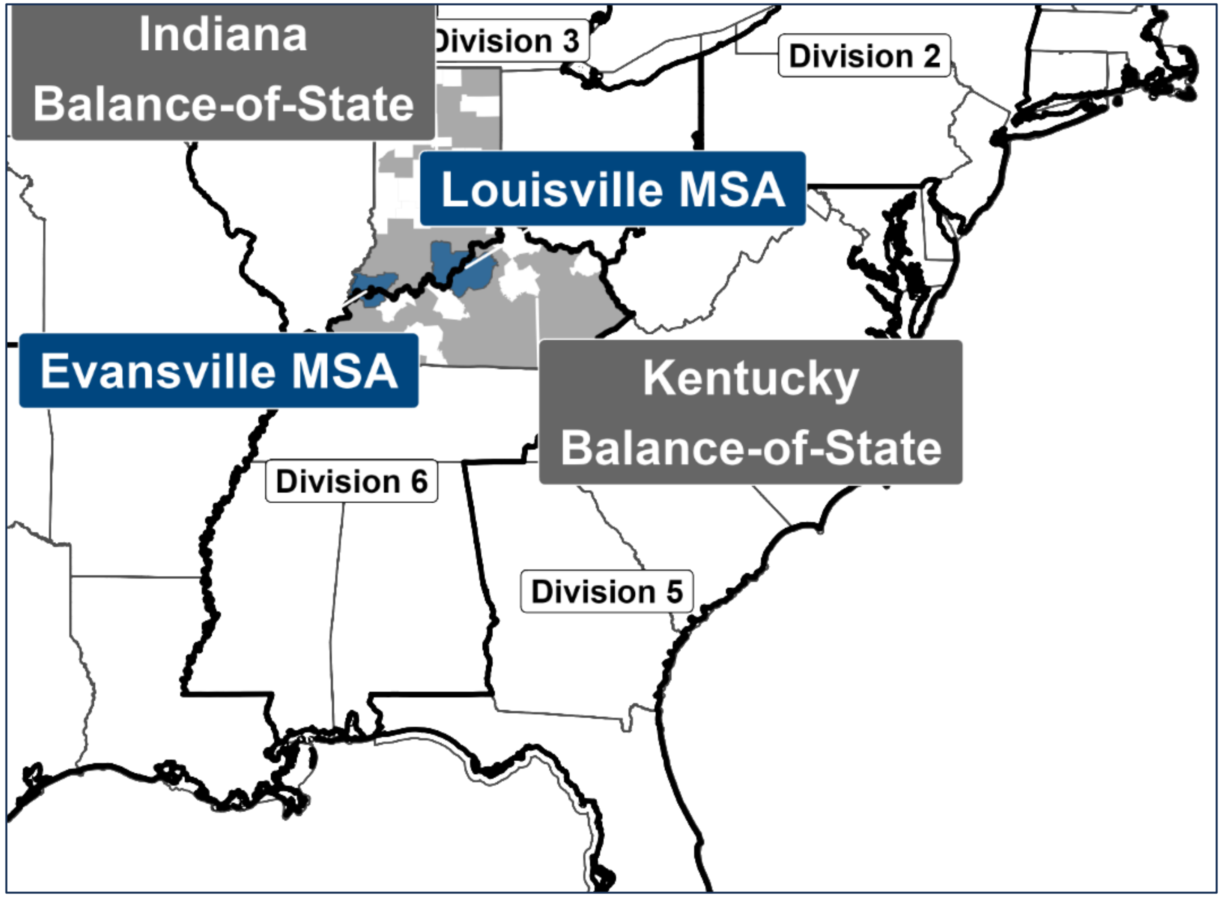
- Six-digit standard occupational classification codes (SOC6)
  - SOC2: 15-0000

    “Computer and Mathematical Occupations
  - SOC4: 15-2000

    “Mathematical Science Occupations
  - SOC6: 15-2041

    “Statisticians”

Geographies: MSA and BOS



Available Data

National Compensation Survey (NCS)

- Measures wages and **measures benefits**
- **Small sample size** with limited utility for subnational estimates

Occupational Employment Statistics (OES) Program*

- Measures wages but doesn't **measure benefits**
- **Large sample size** available for subnational estimates

** Recently renamed to Occupational Employment and Wage Statistics (OEWS) program*

Need for Data Integration: Benefits Only Observed in Some Domains

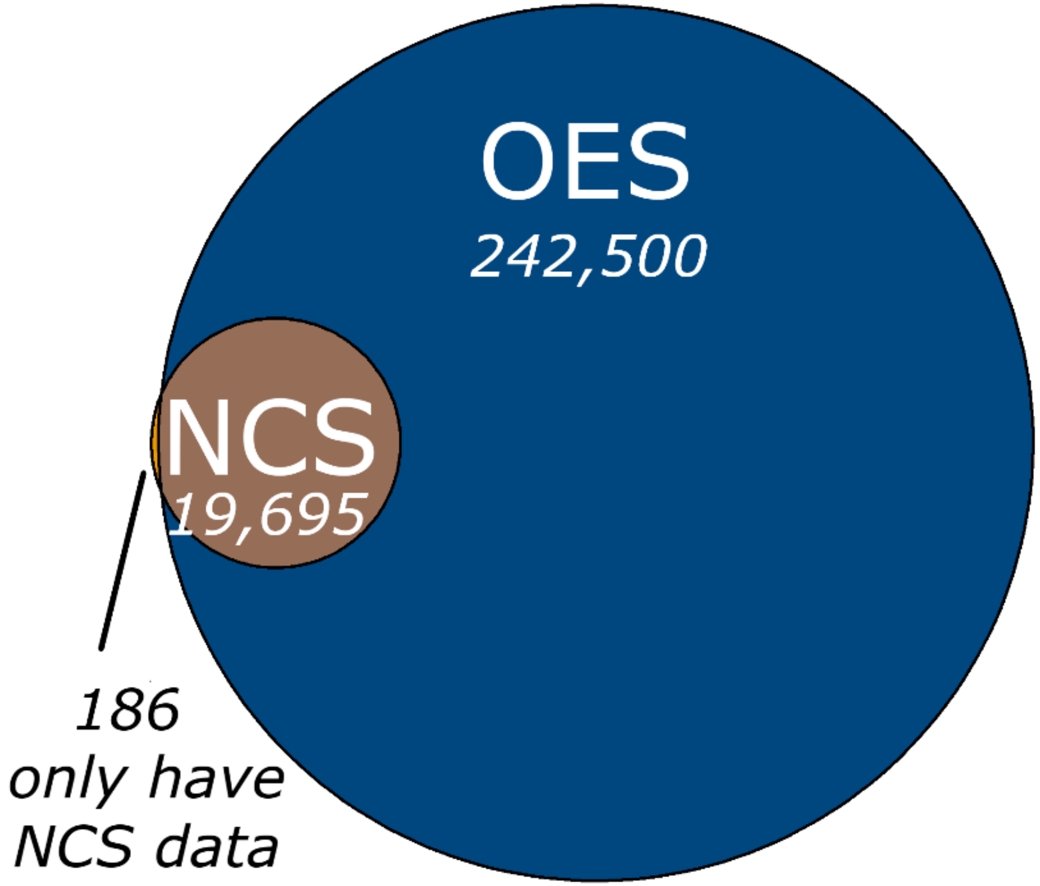
Occupations

- Almost all domains have an OES wage estimate
- Only 8% of domains have a benefits estimate available from NCS
- NCS availability improves at higher levels of aggregation
  - Availability for SOC6 domains:

    63% at Census division level

    94% nationally

MSA/BOS x SOC6 Domains

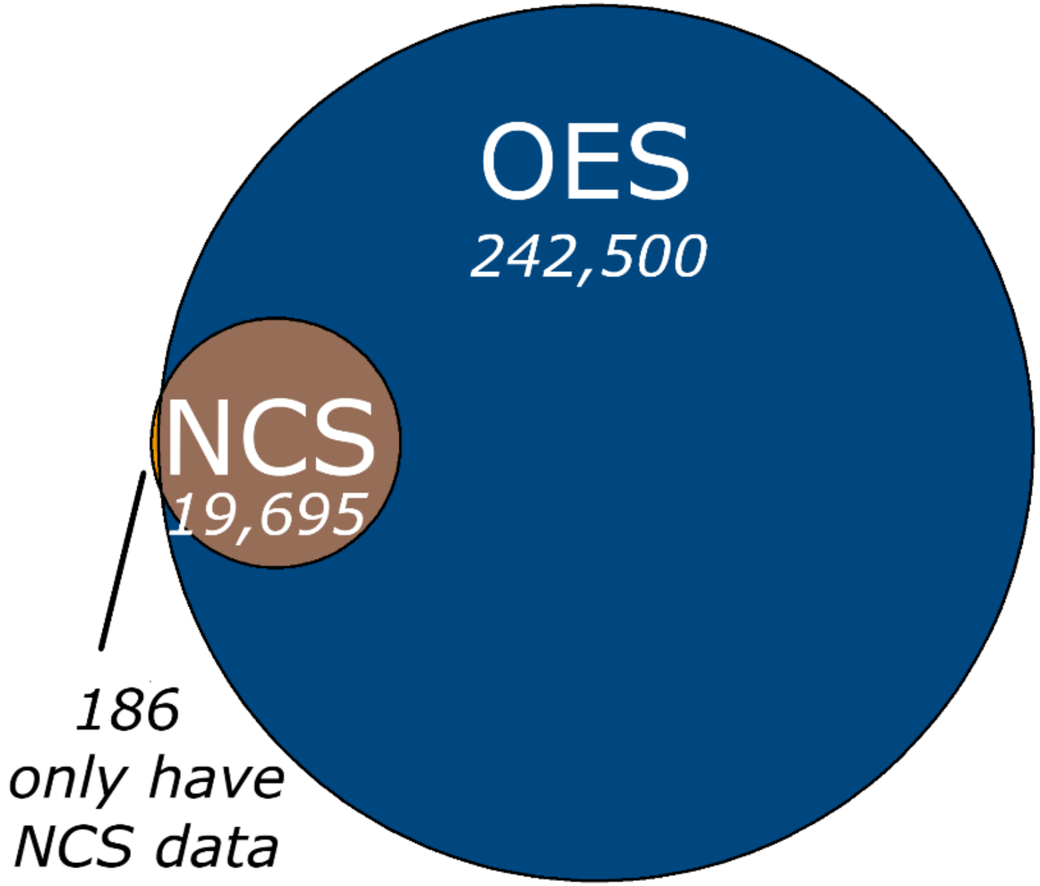


Need for Data Integration: Benefits Only Observed in Some Domains

Occupations

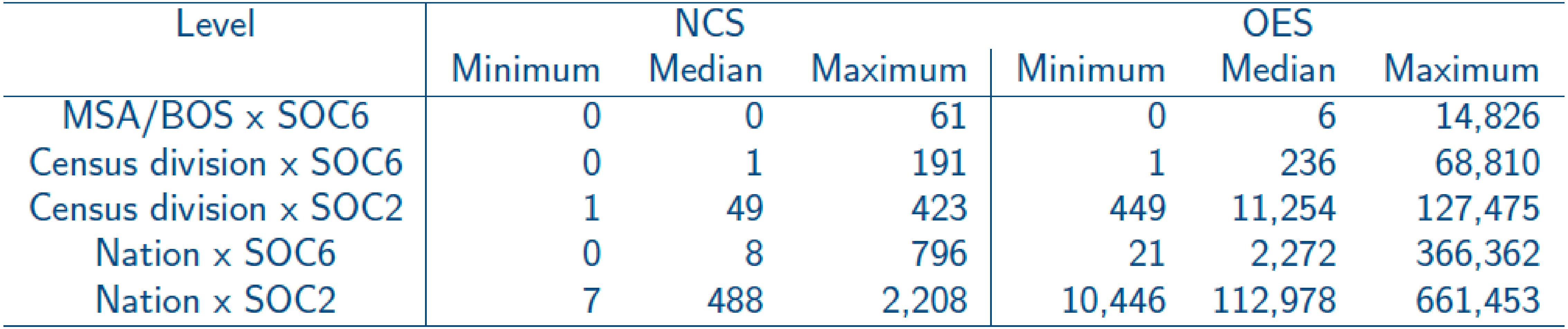
- Almost all domains have an OES wage estimate
- Only 8% of domains have a benefits estimate available from NCS
- NCS availability improves at higher levels of aggregation
  - Availability for SOC2 domains: 100% at Census division level

MSA/BOS x SOC6 Domains



Challenge: Small Sample Sizes

Summary of sample sizes of domains, 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 among domains with any NCS observations
- Median OES sample size is 5 in OES-only domains and 6 in all OES domains

- Currently, BLS doesn’t publish total compensation estimates for occupations within geographies

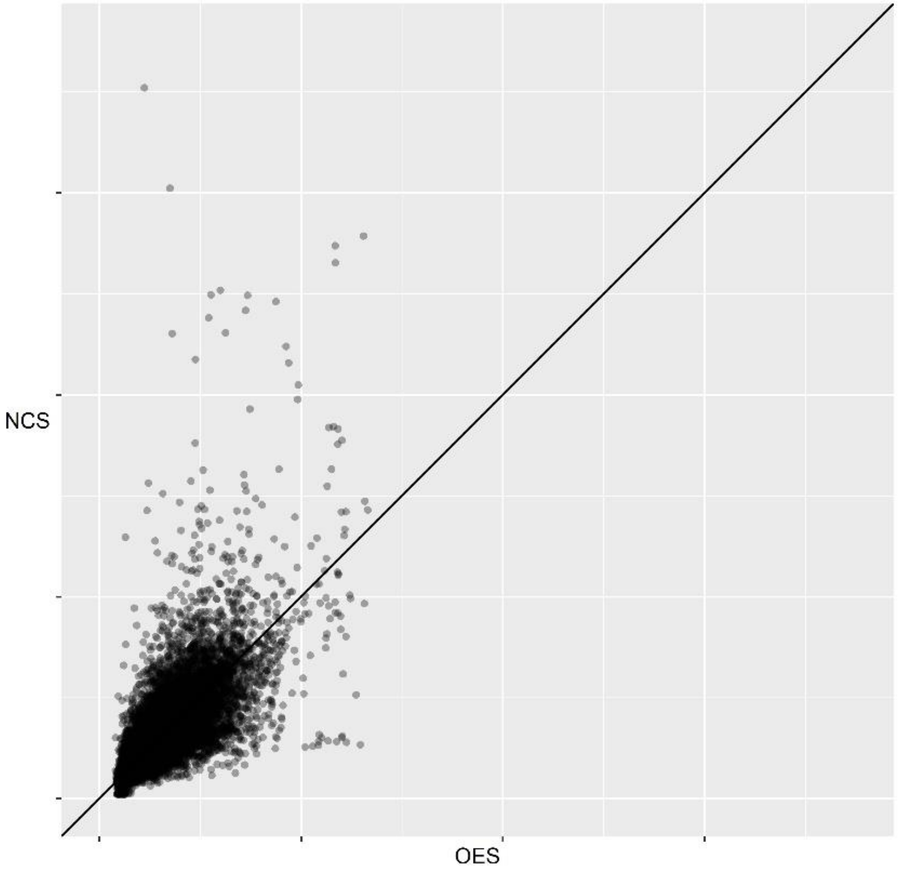
Need for Data Integration: Benefits Only Observed in Some Domains

Occupations

- The NCS wage estimate for a given domain is typically noisier than the OES estimate
- Presence of different estimates from different surveys may be confusing for data users

** Visualization Note: Two (large) domain-level NCS wage estimates removed to improve visualization*

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



Modeling Estimation Approach

- Data integration using bivariate modeling:
  - Use the strong relationship between wages and benefits to predict benefits for domains that only have an OES wage estimate
  - Combine OES and NCS wage estimates into a single wage estimate
- Smoothing through small area estimation methods:
  - Improve upon the precision of the OES and NCS direct estimates by borrowing strength across domains
- Hierarchical Bayesian modeling to accomplish both goals
  - Treat unknown population wage and benefits as bivariate latent characteristic
  - Fay-Herriot type model with sampling level and smoothing levels
  - Exploit nested structure of domains when smoothing

Modeling Inputs

- Wage and benefit estimates in $/hr, on log scale
- Accompanying variance estimates
  - Stabilized using the projection method of Erciulescu & Opsomer (2019)
- Covariates $\boldsymbol{x_i}$ defined in terms of area type (MSA or BOS), census division, and two-way interactions
- Identifiers for SOC2 and SOC6

Survey	Point Estimate	Variance Estimate
OES	y_i^{OES}	$\left(\sigma_{1,i}^{OES}\right)^2$
NCS	$y_i^{NCS} = (y_{1,i}^{NCS}, y_{2,i}^{NCS})$	Σ_i^{NCS}

Bivariate Hierarchical Bayes Multi-fold Model

Sampling Level

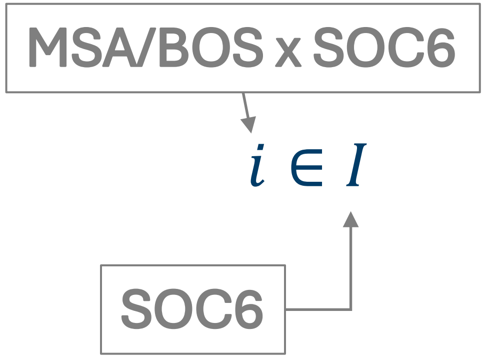
$$y_i^{NCS} \sim \mathrm{N}\left(\theta_i, \Sigma_i^{NCS}\right) \qquad i \text{ observed in NCS}$$

$$y_{1,i}^{OES} \sim \mathrm{N}\left(\theta_{1,i}, \left(\sigma_i^{OES}\right)^2\right) \qquad i \text{ observed in OES}$$

Smoothing Level

$$\theta_i \sim N(x_i'\beta + u_I, \Sigma_b), i \text{ observed in NCS or OES}, \qquad i \in I$$

$$u_I \sim N(0, \Sigma_u)$$



Prior Distributions

$$\beta \sim \mathrm{N}(0, 10^4) \qquad \text{component–wise}$$

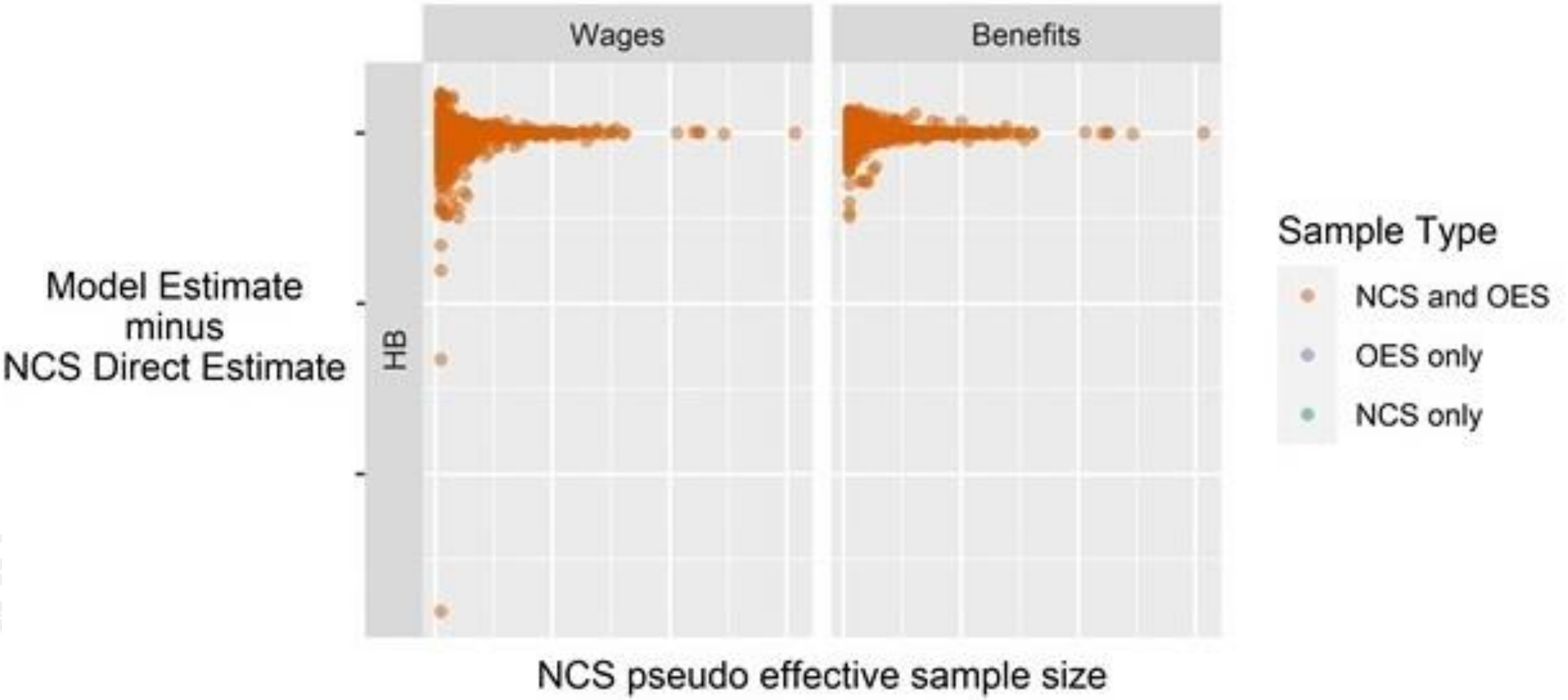
$$(\Sigma_b, \Sigma_u) \sim \text{Inverse–Wishart}\,(I_2, 3) \qquad \text{component–wise}$$

Model Fit, Assumption Checks, Predictions

- Fit
  - R JAGS
  - Markov chain Monte Carlo (MCMC): 2,100 samples for inference
    - 3 chains with 10,000 samples each. 3,000 burn-in, 1-in-10 thinning to reduce storage
  - Models fit separately for the 22 major occupation groups (SOC2)
- Assumption checks
  - MCMC diagnostics: $\hat{R}$, MC effective sample size, MC standard error, autocorrelation
  - Model specification: posterior predictive checks
- Prediction
  - Marginal posterior distribution for θ_i
  - Transformations: exponential, sum

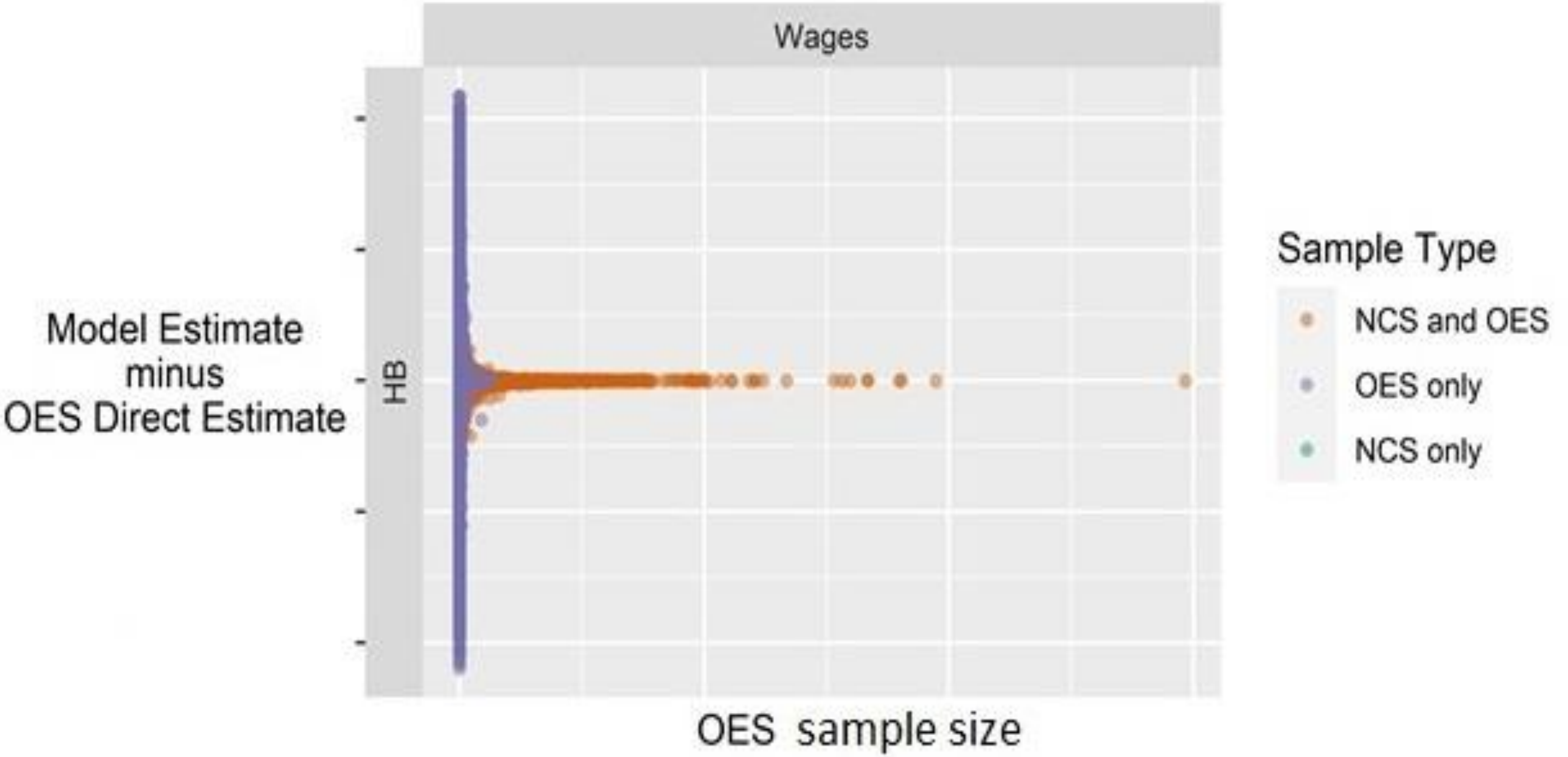
Comparison of NCS and Model: Point Estimates

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

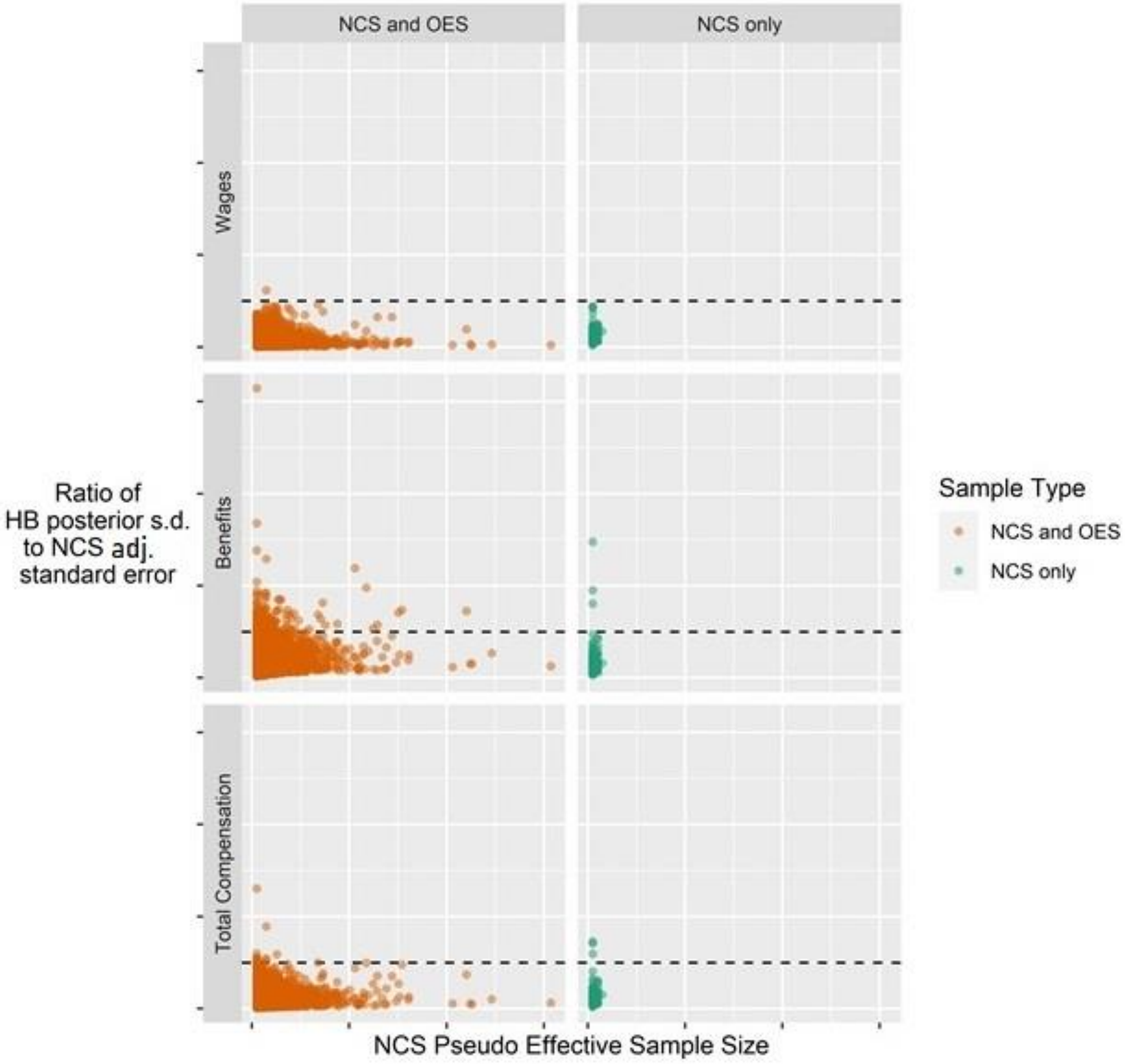


Comparison of OES and Model: Point Estimates

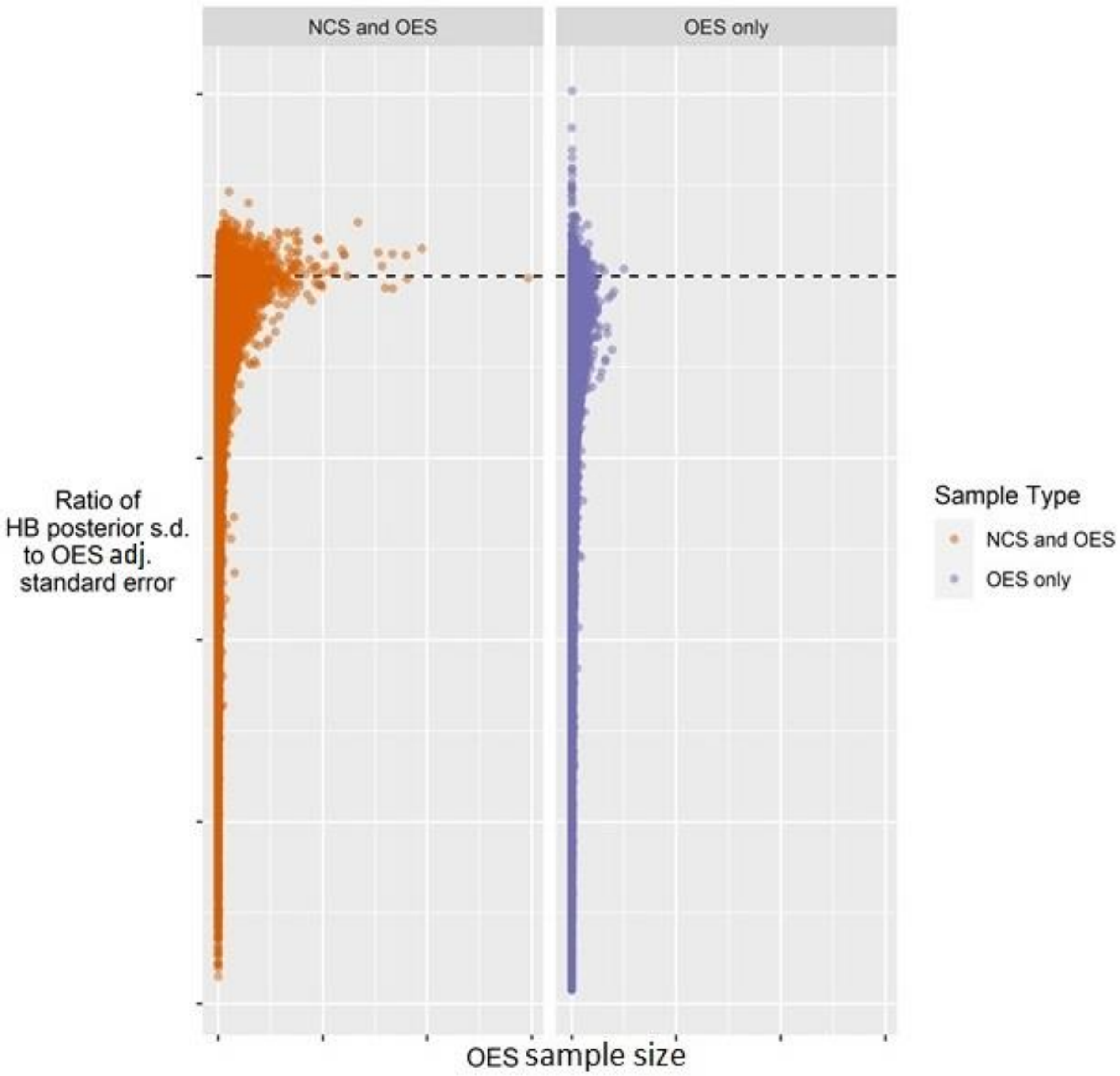
Domain-level wage and benefits estimates, MSA/BOS x SOC6



Comparison of NCS and Model: Standard Errors

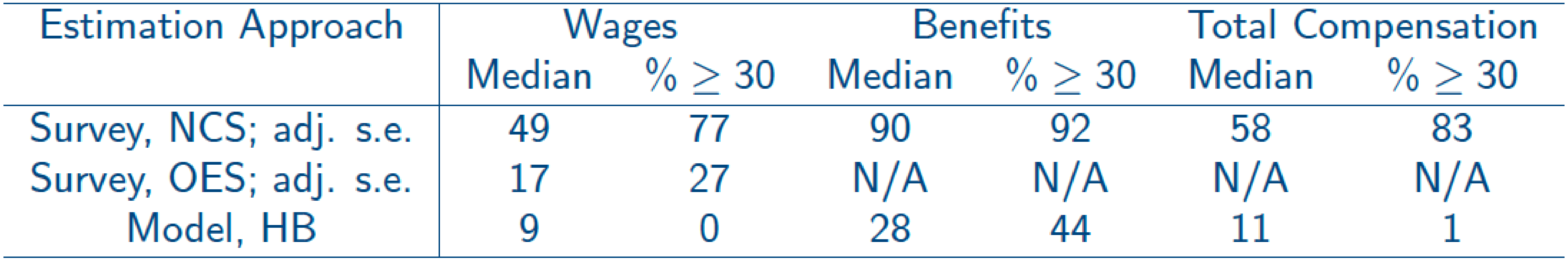


Comparison of OES and Model: Standard Errors



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

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



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

- Application results in a complete set of wage, benefits, and total compensation estimates for all domains of interest, with associated uncertainty measures
- Hierarchical model estimates have improved precision compared to direct estimates from the survey
- Applicable to other scenarios where two surveys collect information on a common variable that is strongly related to a variable only measured in one survey

Full Paper:

Erciulescu, A.L., Opsomer, J.D. and Schneider, B.J. (2023). “Statistical data integration using multilevel models to predict employee compensation.” Canadian Journal of Statistics, 51: 312-326.

https://doi.org/10.1002/cjs.11688

Comments, Questions, Suggestions?

BenjaminSchneider@westat.com