Discussion of Semiparametric Bayesian Density Estimation with Disparate Data Sources

Paper by: Finucane, Paciorek, Stevens and Ezzati,
Journal of the American Statistics Association,
to appear 2015.

Daniell Toth

U.S. Bureau of Labor Statistics

Office of Survey Methods Research

Content represents only the opinion of the author.

Outline of Discussion

(1) Brief review of method

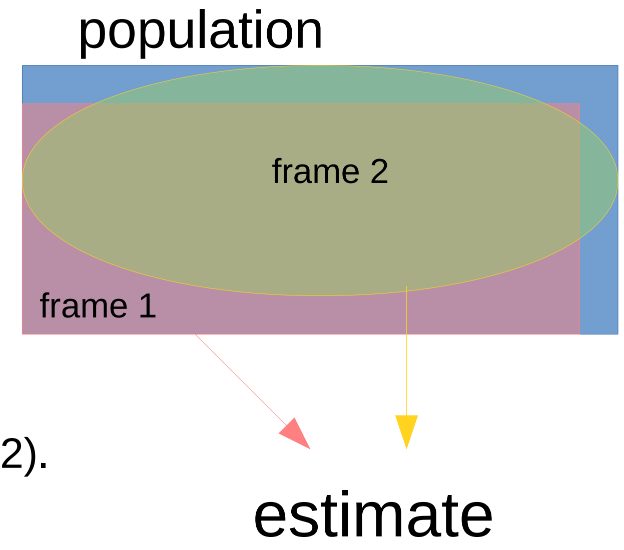
(2) Suggest an application to BLS economic data

A Common Goal

Combine data from several sources to produce a single estimate.

may have

- different sample design
- different frame
- different coverage



Dong, Qi. "Combining Information from Multiple Complex Surveys." (2012).

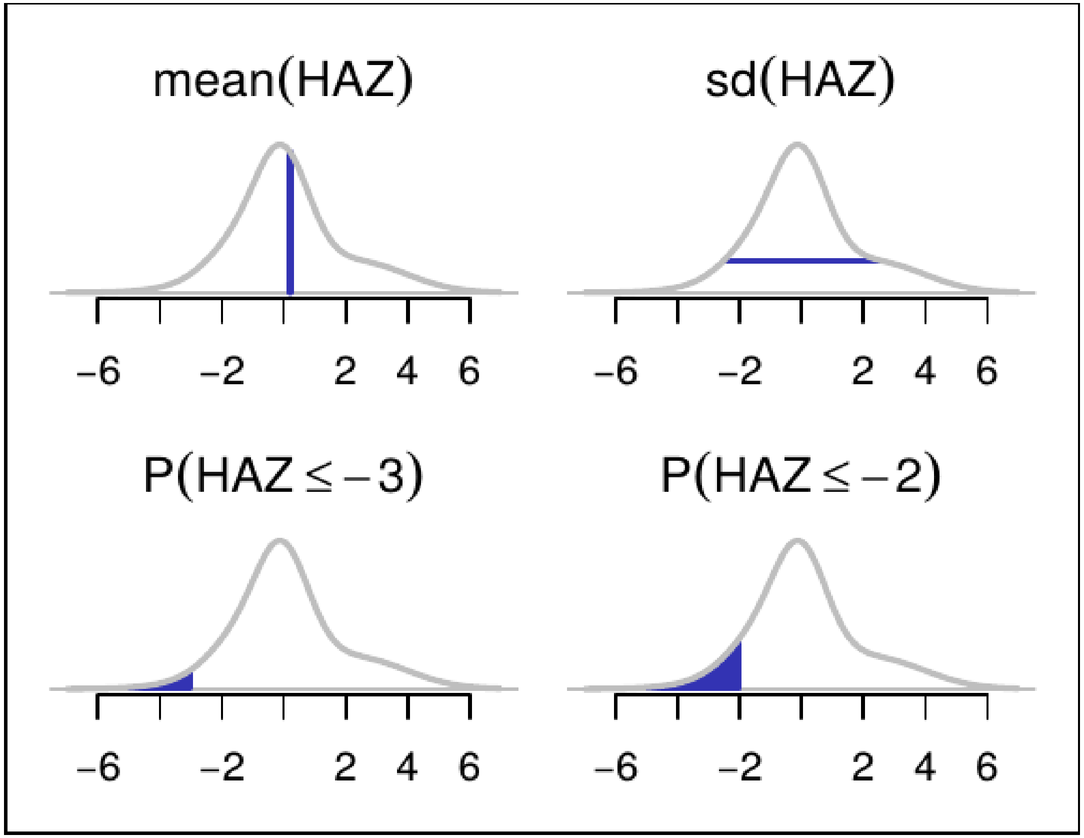
Hentschel, Jesko, et al. "Combining census and survey data to trace the spatial dimensions of poverty: A case study of Ecuador." The World Bank Economic Review 14.1 (2000): 147-165.

Lohr, Sharon L., and J. Michael Brick. "Blending domain estimates from two victimization surveys with possible bias." Canadian Journal of Statistics 40.4 (2012): 679-696.

Lohr, Sharon, and JN K. Rao. "Estimation in multiple-frame surveys." Journal of the American Statistical Association 101.475 (2006): 1019-1030.

Interested in the Entire Distribution

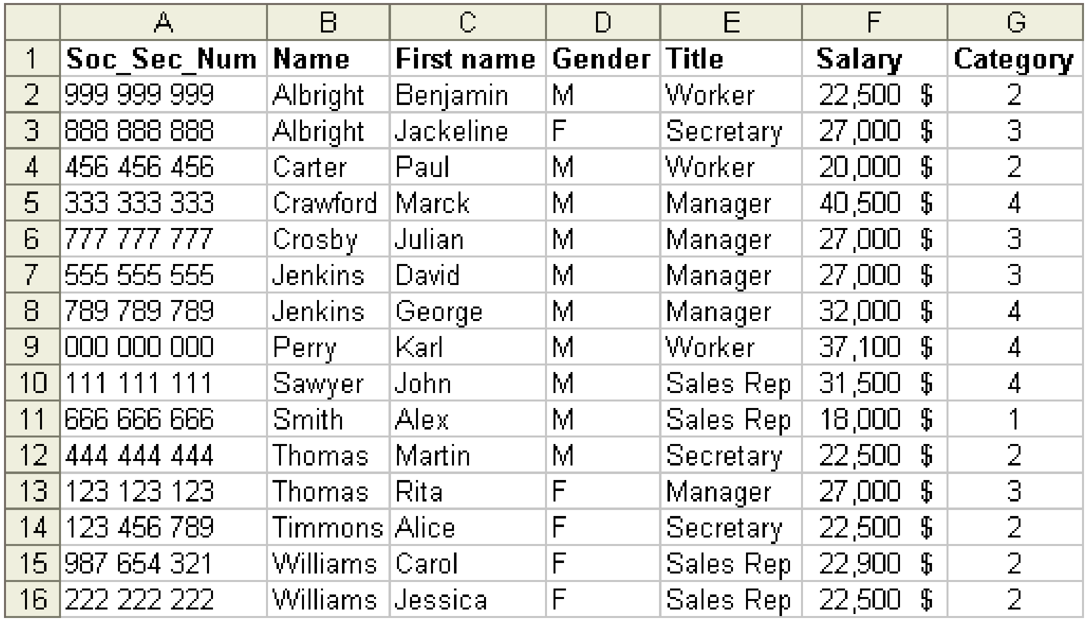
Estimate functionals of the underlying population distribution from several sources of sample data.



Finucane, Paciorek, Stevens, and Ezzati; *JASA*, in press

Uses Summary Statistics

individual unit data



	A	B	C	D	E	F	G
1	**Soc_Sec_Num**	**Name**	**First name**	**Gender**	**Title**	**Salary**	**Category**
2	999 999 999	Albright	Benjamin	M	Worker	22,500 $	2
3	888 888 888	Albright	Jackeline	F	Secretary	27,000 $	3
4	456 456 456	Carter	Paul	M	Worker	20,000 $	2
5	333 333 333	Crawford	Marck	M	Manager	40,500 $	4
6	777 777 777	Crosby	Julian	M	Manager	27,000 $	3
7	555 555 555	Jenkins	David	M	Manager	27,000 $	3
8	789 789 789	Jenkins	George	M	Manager	32,000 $	4
9	000 000 000	Perry	Karl	M	Worker	37,100 $	4
10	111 111 111	Sawyer	John	M	Sales Rep	31,500 $	4
11	666 666 666	Smith	Alex	M	Sales Rep	18,000 $	1
12	444 444 444	Thomas	Martin	M	Secretary	22,500 $	2
13	123 123 123	Thomas	Rita	F	Manager	27,000 $	3
14	123 456 789	Timmons	Alice	F	Secretary	22,500 $	2
15	987 654 321	Williams	Carol	F	Sales Rep	22,900 $	2
16	222 222 222	Williams	Jessica	F	Sales Rep	22,500 $	2

summary statistics

mean salary

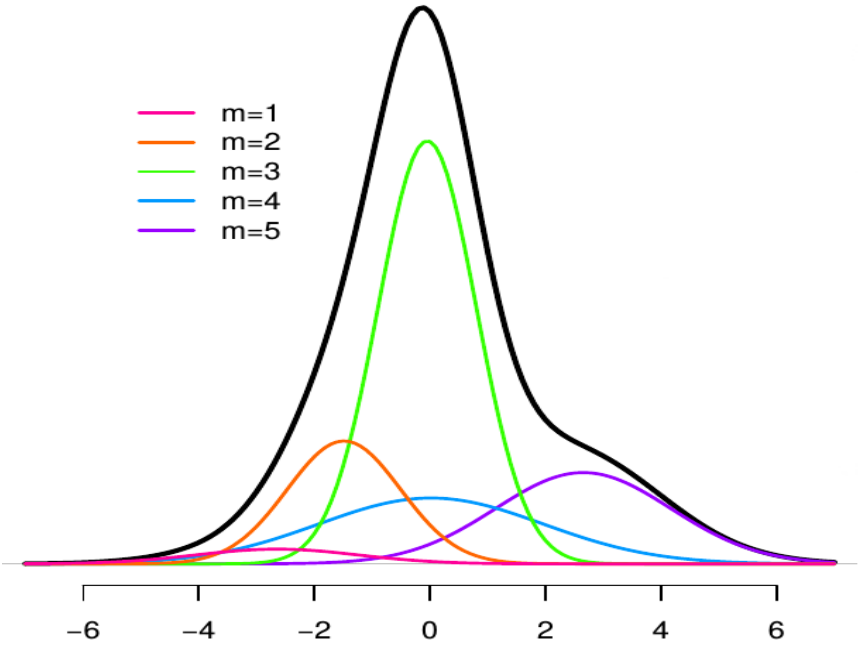
median salary

salary of top 1%

$f(y)$ estimate

Simple Likelihood

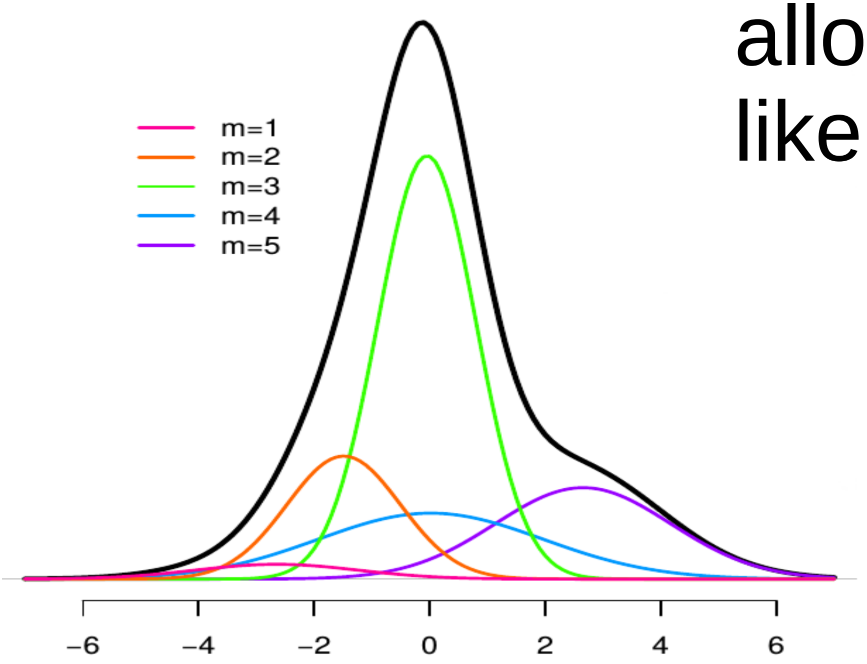
$$f_i(y) = \sum w_{mi} \, \mathcal{N}(\mu_m, \sigma_m^2)$$



Finucane, Paciorek, Stevens, and Ezzati; *JASA*, in press

Simple Likelihood

$$f_i(y) = \sum w_{mi}\, \mathcal{N}(\mu_m, \sigma_m^2)$$



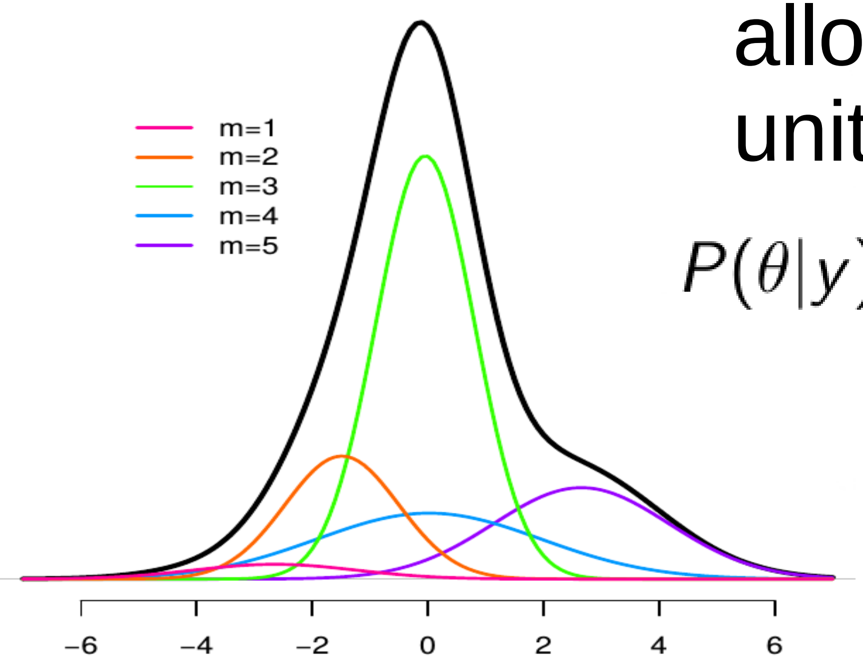
allows easy calculation of likelihood for summary statistics

$$P(y_{\text{summary}}|\theta)$$

Finucane, Paciorek, Stevens, and Ezzati; *JASA*, in press

Simple Likelihood

$$f_i(y) = \sum w_{mi}\, \mathcal{N}(\mu_m, \sigma_m^2)$$



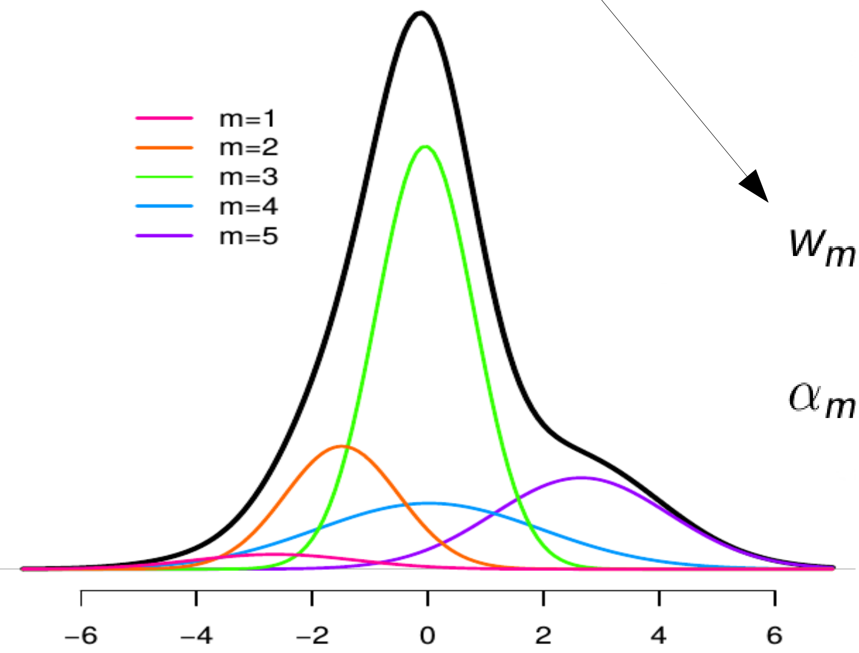
allows combining with individual unit data

$$P(\theta|y) \propto P(y_{\text{individual}}|\theta)\ P(y_{\text{summary}}|\theta)\ P(\theta)$$

Finucane, Paciorek, Stevens, and Ezzati; *JASA*, in press

Complicated Weights

$$f_i(y) = \sum w_{mi}\, \mathcal{N}(\mu_m, \sigma^2_m)$$



$$w_{mi} = \Phi(\alpha_{mi}) \prod_{u=1}^{m-1} (1 - \Phi(\alpha_{ui}))$$

$$\alpha_{mi} \sim \mathcal{N}(a^c_{mj[i]} + b^c_{mj[i]} t_i + u_{mj[i],t_i} + X_i \beta_m + e_{mi}, \tau^2_{mi})$$

Finucane, Paciorek, Stevens, and Ezzati; *JASA*, in press

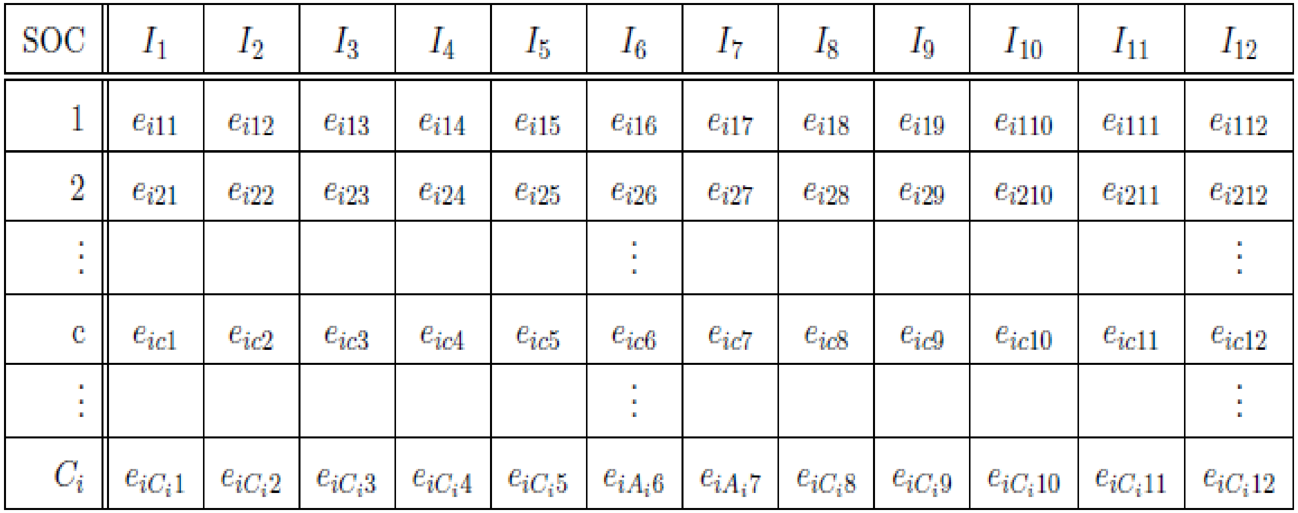
Potential Application
BLS Occupational Wage Data

Occupational Employment Statistics Survey (OES)

- Semi-annual establishment survey (May and Nov)
- PPS Stratified Sample of establishments
- Sample size of about 179,000 establishments
- Measures employment and wages by occupation
- 78% response rate

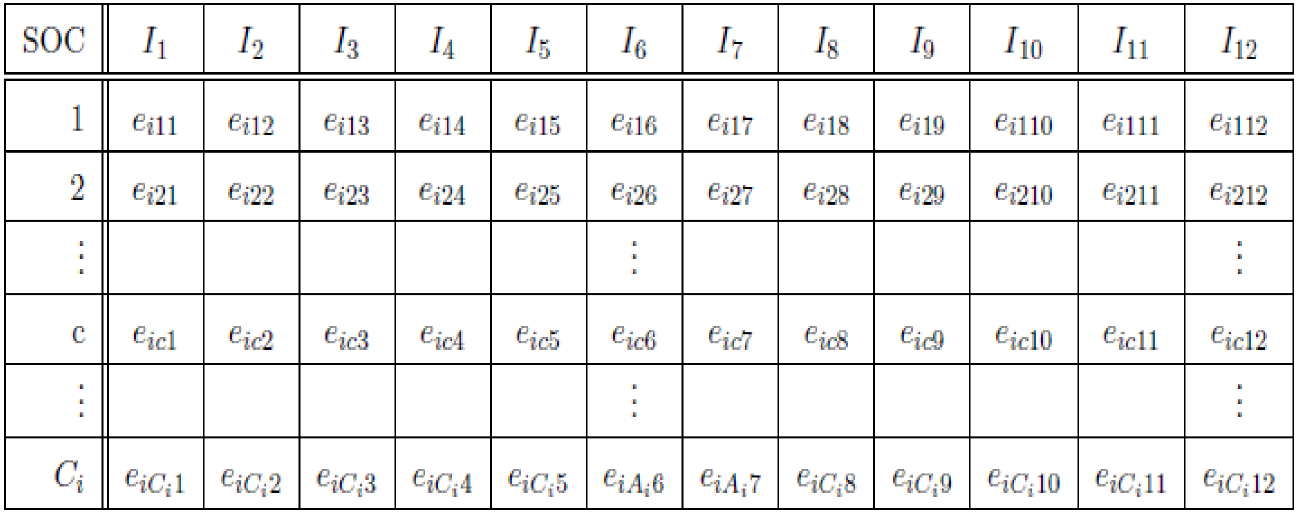
OES publishes employment and wage rate estimates for 800 occupations by industry and area.

OES Data Comes as Cell Counts



SOC	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}
1	e_{i11}	e_{i12}	e_{i13}	e_{i14}	e_{i15}	e_{i16}	e_{i17}	e_{i18}	e_{i19}	e_{i110}	e_{i111}	e_{i112}
2	e_{i21}	e_{i22}	e_{i23}	e_{i24}	e_{i25}	e_{i26}	e_{i27}	e_{i28}	e_{i29}	e_{i210}	e_{i211}	e_{i212}
$\vdots$						$\vdots$						$\vdots$
c	e_{ic1}	e_{ic2}	e_{ic3}	e_{ic4}	e_{ic5}	e_{ic6}	e_{ic7}	e_{ic8}	e_{ic9}	e_{ic10}	e_{ic11}	e_{ic12}
$\vdots$						$\vdots$						$\vdots$
C_i	e_{iC_i1}	e_{iC_i2}	e_{iC_i3}	e_{iC_i4}	e_{iC_i5}	e_{iA_i6}	e_{iA_i7}	e_{iC_i8}	e_{iC_i9}	e_{iC_i10}	e_{iC_i11}	e_{iC_i12}

OES Data Comes as Cell Counts



SOC	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}
1	e_{i11}	e_{i12}	e_{i13}	e_{i14}	e_{i15}	e_{i16}	e_{i17}	e_{i18}	e_{i19}	e_{i110}	e_{i111}	e_{i112}
2	e_{i21}	e_{i22}	e_{i23}	e_{i24}	e_{i25}	e_{i26}	e_{i27}	e_{i28}	e_{i29}	e_{i210}	e_{i211}	e_{i212}
$\vdots$						$\vdots$						$\vdots$
c	e_{ic1}	e_{ic2}	e_{ic3}	e_{ic4}	e_{ic5}	e_{ic6}	e_{ic7}	e_{ic8}	e_{ic9}	e_{ic10}	e_{ic11}	e_{ic12}
$\vdots$						$\vdots$						$\vdots$
C_i	e_{iC_i1}	e_{iC_i2}	e_{iC_i3}	e_{iC_i4}	e_{iC_i5}	e_{iA_i6}	e_{iA_i7}	e_{iC_i8}	e_{iC_i9}	e_{iC_i10}	e_{iC_i11}	e_{iC_i12}

Some establishments voluntarily provide **full salary data** for every employee.

QCEW Data

We also have quarterly administrative payroll records for almost every establishment through the Unemployment Insurance records.

Contains: location; industry; total number of employees; total wages paid.

We can compute an approximate average wage

total quarterly payroll of establishment divided by total employment

AVEWAGE = WAGE/EMPL

QCEW Data

We also have quarterly administrative payroll records for almost every establishment through the Unemployment Insurance records.

Contains: location; industry; total number of employees; total wages paid.

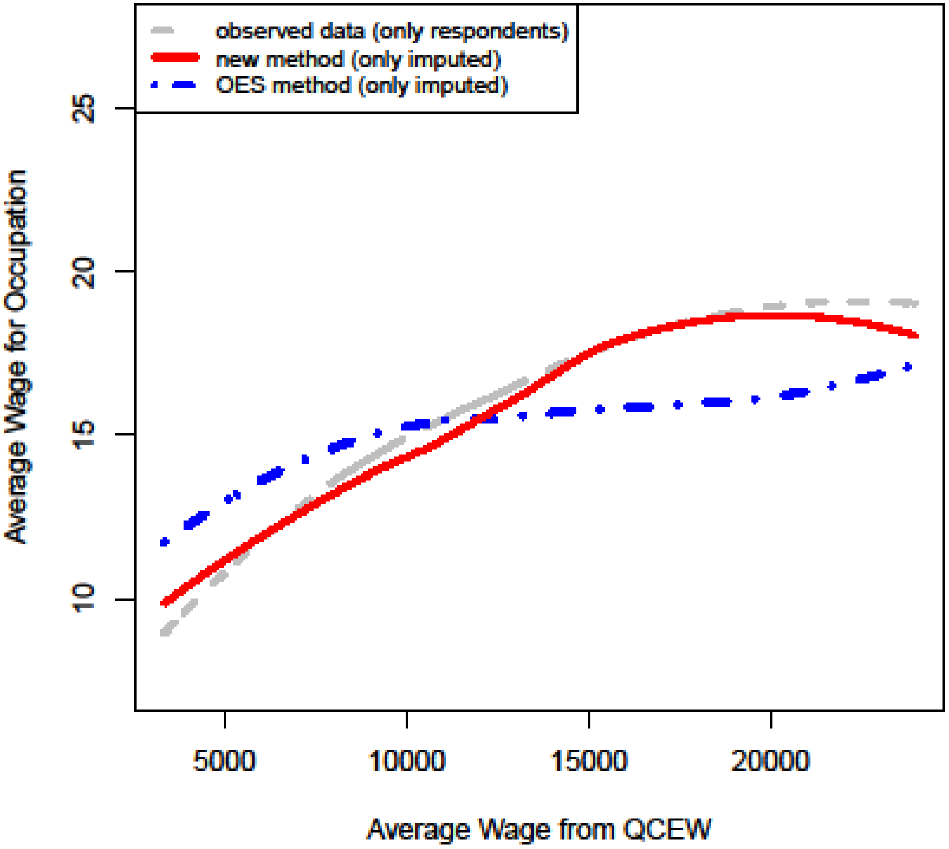
We can compute an approximate average wage

total quarterly payroll of establishment divided by total employment

AVEWAGE = WAGE/EMPL

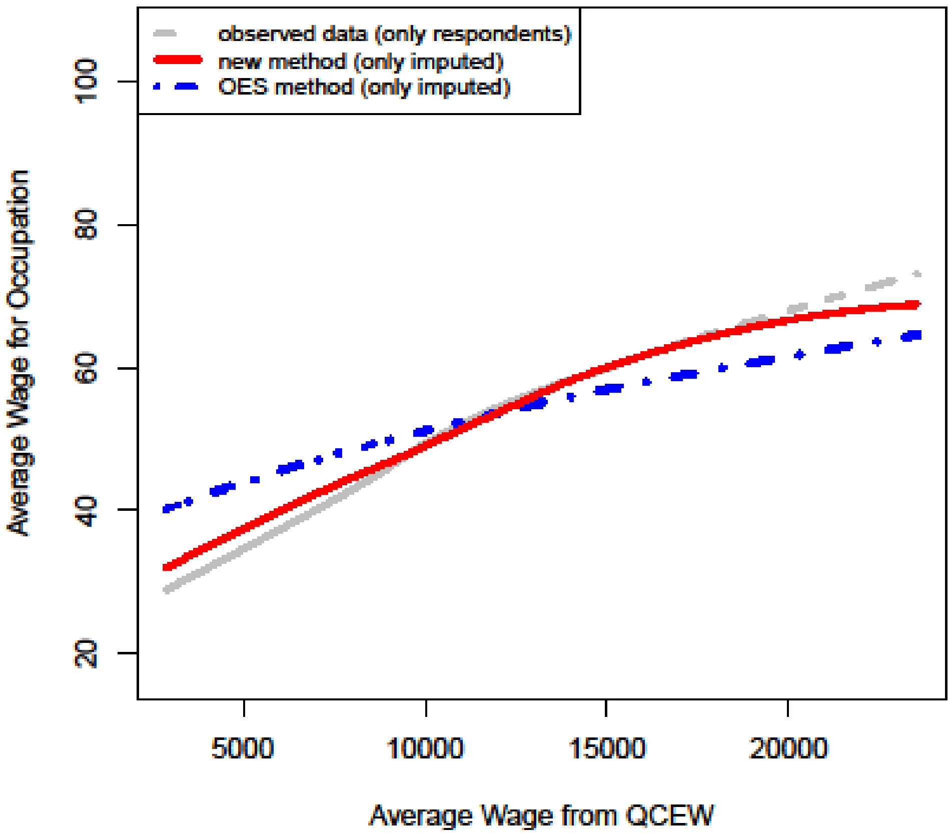
This does not include occupational information or hours worked.

Customer Service Rep



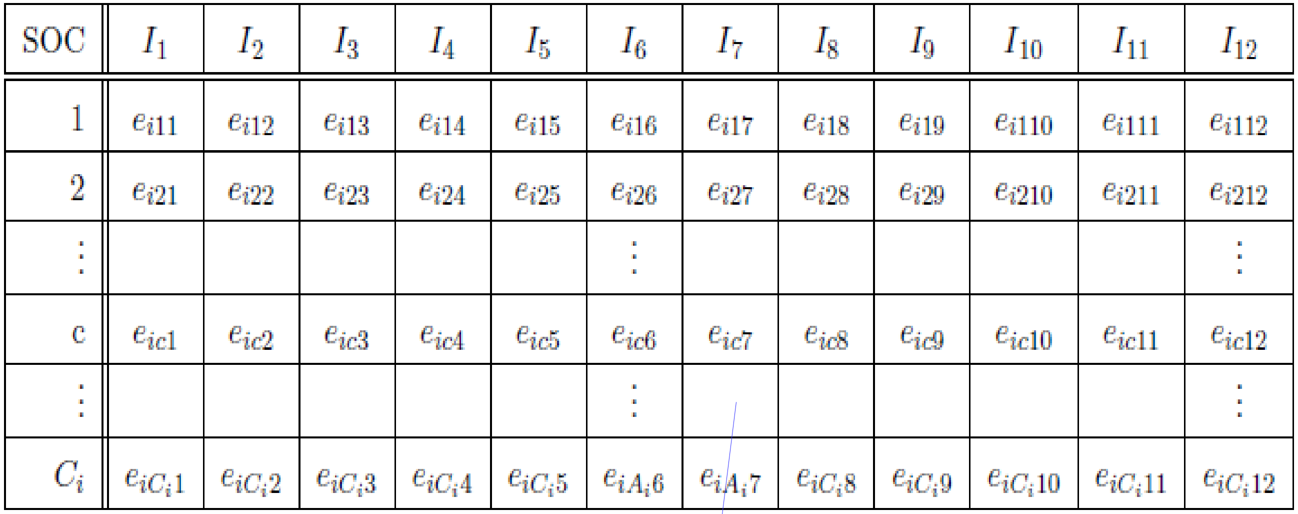
Horton, Toth and Phipps (2014), *Annals of Applied Statistics*, 8, 956-973

Managers



Horton, Toth and Phipps (2014), *Annals of Applied Statistics*, 8, 956-973

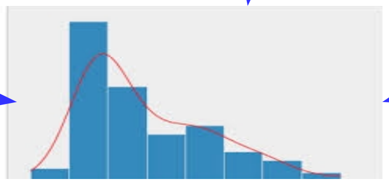
Estimation with Disparate Data Sources



SOC	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}
1	e_{i11}	e_{i12}	e_{i13}	e_{i14}	e_{i15}	e_{i16}	e_{i17}	e_{i18}	e_{i19}	e_{i110}	e_{i111}	e_{i112}
2	e_{i21}	e_{i22}	e_{i23}	e_{i24}	e_{i25}	e_{i26}	e_{i27}	e_{i28}	e_{i29}	e_{i210}	e_{i211}	e_{i212}
$\vdots$						$\vdots$						$\vdots$
c	e_{ic1}	e_{ic2}	e_{ic3}	e_{ic4}	e_{ic5}	e_{ic6}	e_{ic7}	e_{ic8}	e_{ic9}	e_{ic10}	e_{ic11}	e_{ic12}
$\vdots$						$\vdots$						$\vdots$
C_i	e_{iC_i1}	e_{iC_i2}	e_{iC_i3}	e_{iC_i4}	e_{iC_i5}	e_{iA_i6}	e_{iA_i7}	e_{iC_i8}	e_{iC_i9}	e_{iC_i10}	e_{iC_i11}	e_{iC_i12}

AVEWAGE

Full data from volunteers



Can we estimate underlying distribution?

Thank You

toth.daniell@bls.gov