Training Programme on Statistical Analysis of Disaggregated SDG Indicators for Inclusive Development Policies
Advanced Small Area Estimtion

Partha Lahiri

Joint Program in Survey Methodology & Department of Mathematics
University of Maryland, College Park

UN-SIAP, Chiba, Japan, June 12-13, 2019.

Introduction

Sustainable Development Goals (SDG)

UN-SDG WEB Banner

Small Area (Domain)

A subpopulation of interest with meager or no survey data.

Examples:

- In the National Socioeconomic Characterization Survey (CASEN), conducted by the Chilean government, there are many municipalities (comunas) with small samples or no sample.
- In the National Health and Nutrition Examination Survey (NHANES), a majority of US states do not have sample.
- In a nationwide survey, some cells obtained by cross-classification of age-group, race, gender even at the national level may not have any sample.

Other terms used: local area, sub-domain, small subgroup, subprovince, minor domains.

A classification of domains

- **Planned domains:** These are domains for which separate samples have been planned, designed, and selected.

  Example: States are planned domains if states are treated as design strata.

- **Unplanned domains:** These domains have not been distinguished in the sample selection.

  Example: Age can be considered unplanned domains if samples are obtained from a frame that does not have any information on age.

Demand of small-area statistics

- Small area estimation has been of interest for a long time, especially among demographers to estimate small area population counts and other characteristics of interest in the post-censal years.
- Small-area statistics were used as early as 11th century England and 17th century Canada. In those days, census, special surveys or administrative records were used to obtain small-area statistics.
- There is an increasing demand for diverse, rich and current statistics for small domains for the planning of reform, welfare and administration in many fields and allocation of federal funds to local governments.

State sample sizes with an epsem sample of 10,000 persons

State	1994 Population (in thousands)	Expected sample size
California	31,431	1,207
Texas	18,378	706
New York	18,169	698
.	.	.
.	.	.
.	.	.
Vermont	580	22
DC	570	22
Wyoming	476	18
Total	260,341	10,000

Simulation from the Australian Beef Farm Data

- Finite population: $N = 431$ farms with 50 or more beef cattles that participated in the 1988 Australian Agricultural and Grazing Industries Survey carried out by the Australian Bureau of Agricultural and Resource Economics.
- y: income from beef (y).
- Simulate $R = 1,000$ independent SRS samples, each of of size n, from the finite population. Consider $n = 10,\ 50$.
- Sample means (Horvitz-Thompson or HT estimates), ratio estimates and their associated variance estimates from several simulated samples are displayed in the box plots and compared with the corresponding true values.
- Sample means and the associated variance estimates, though unbiased, exhibit high variability for $n = 10$. Variability decreases as we increase n.

Figure 1: Box Plots of HT and ratio estimates of mean and associated variance estimates

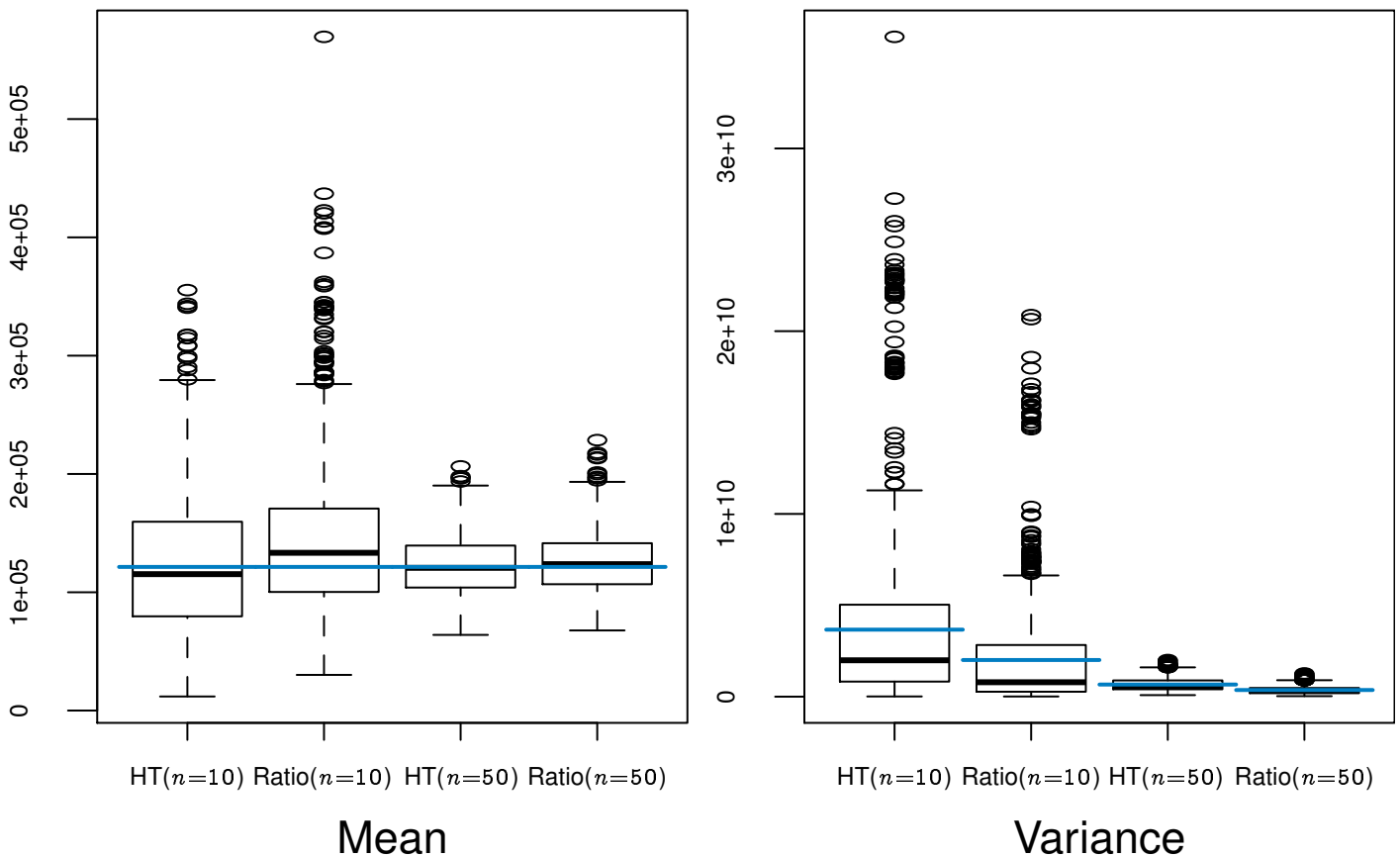


Figure: Time Series Plots of direct poverty rate estimates for selected Comunas

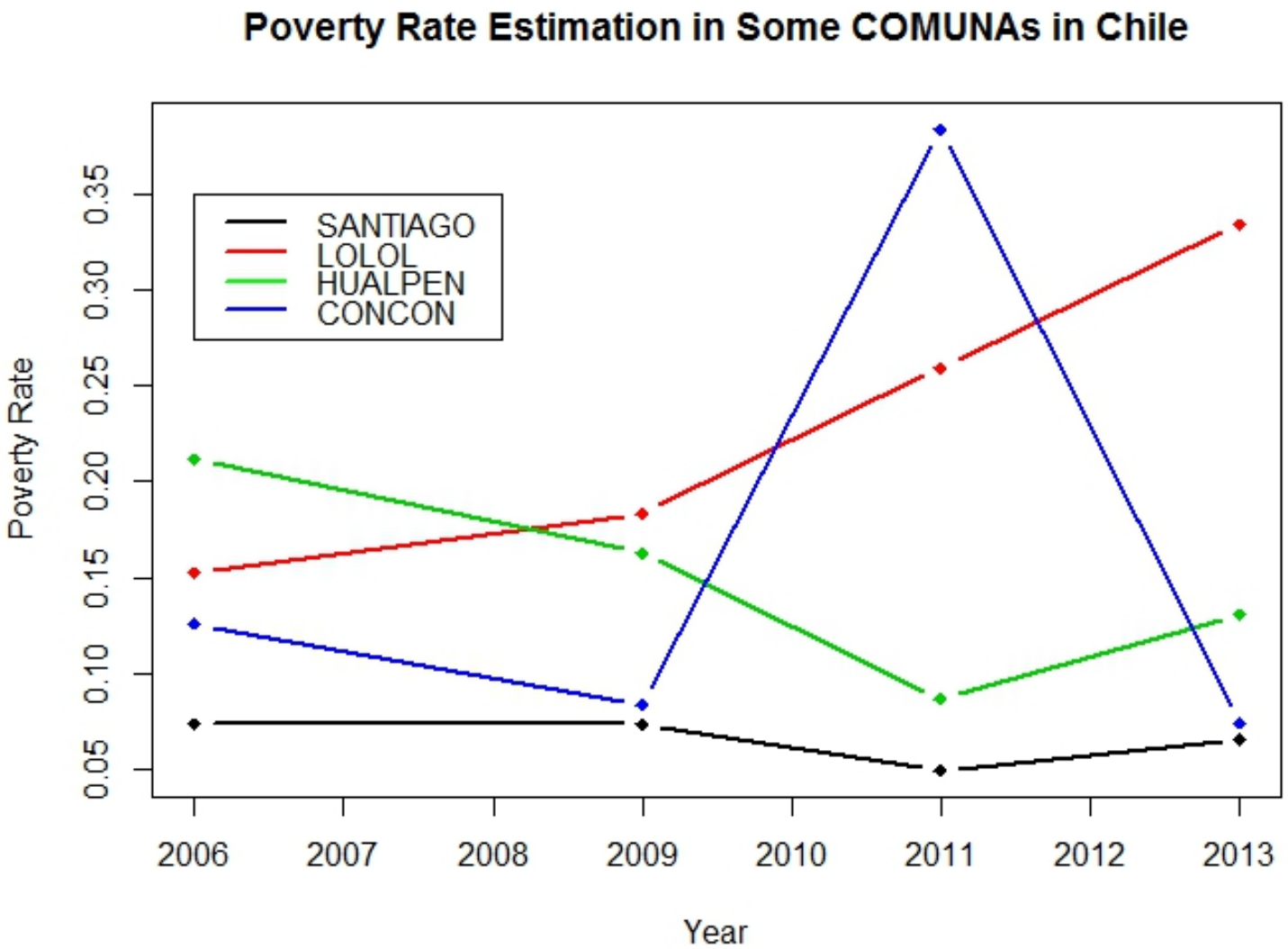
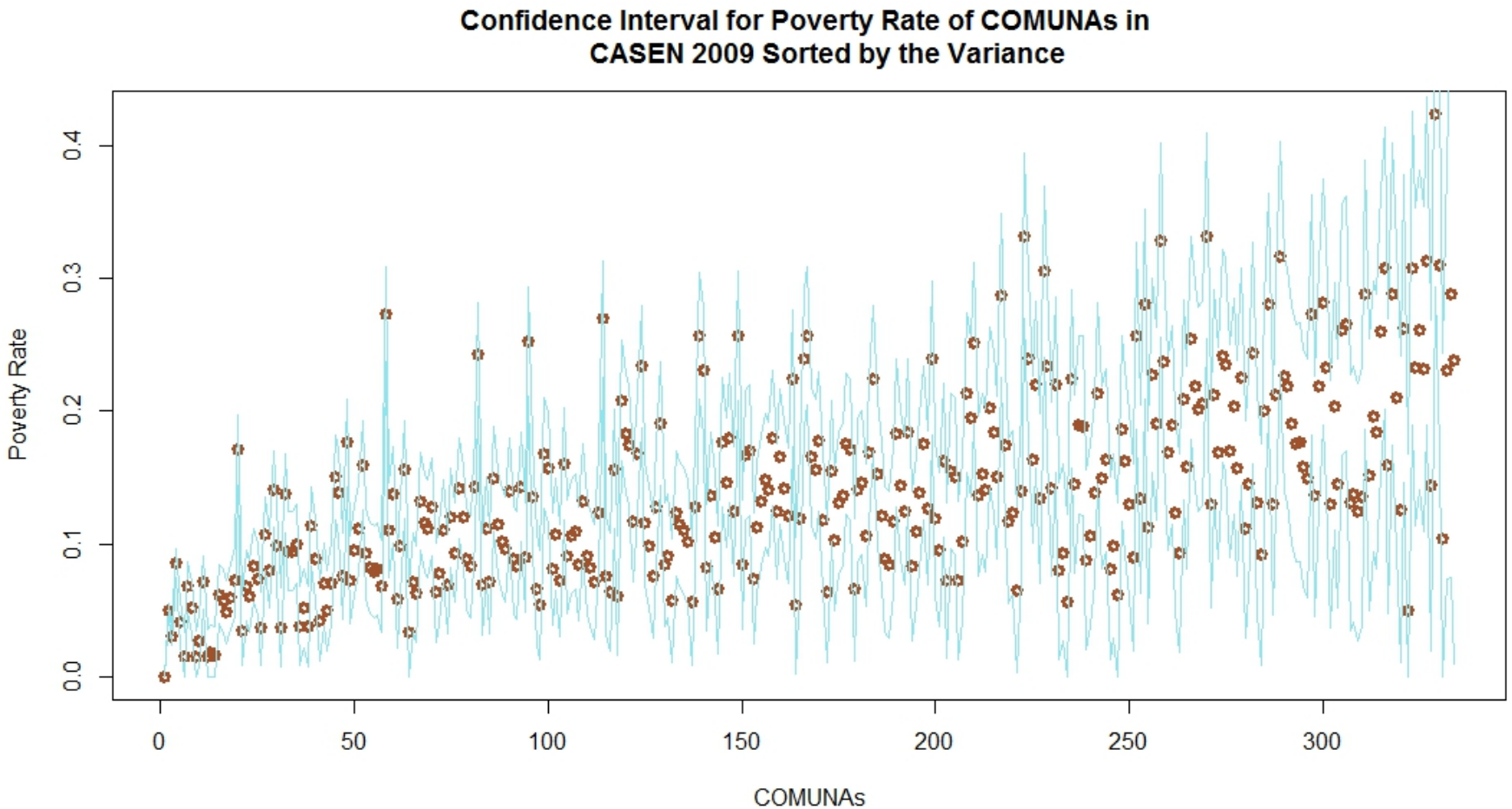


Figure: Direct poverty rate direct estimates and the associated 95% direct confidence intervals for all comunas in CASEN 2009 (sorted by the direct variance estimates)



Alternative data sources

A few examples

Administrative data

Satellite data

Scanner data

Sensor data

GPS data

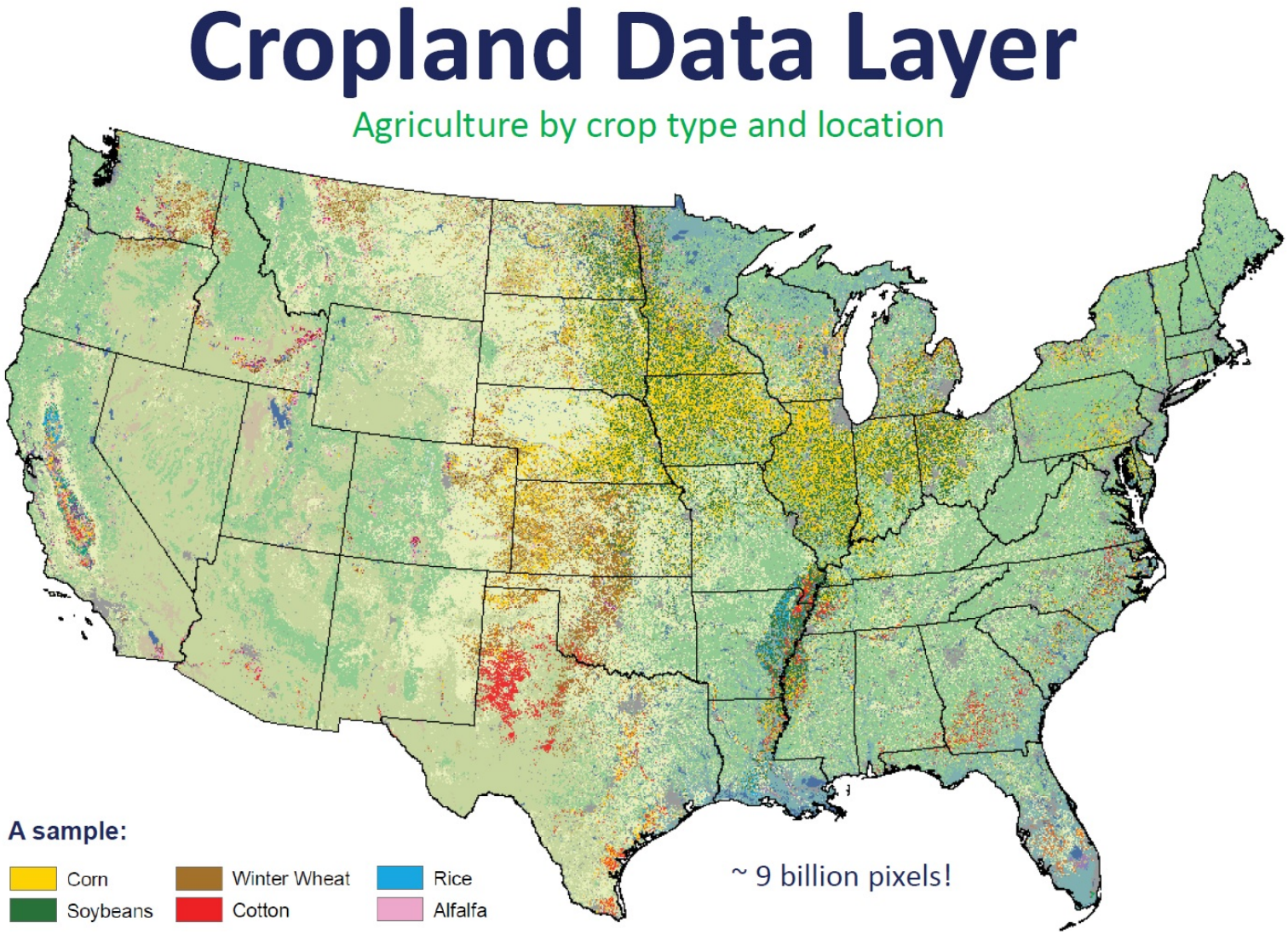
Social media data

Personal data (e.g. data from tracking devices)

Examples of administrative data

- IRS data
- SNAP data
- Insurance enrollment and claims data (e.g. Medicare, Medicaid, BC/BS)
- Hospital discharge data (billing data)

Satellite Data



Zakzeski, A., National Agricultural Statistics Service

Scanner Data: Mango sales in grocery stores

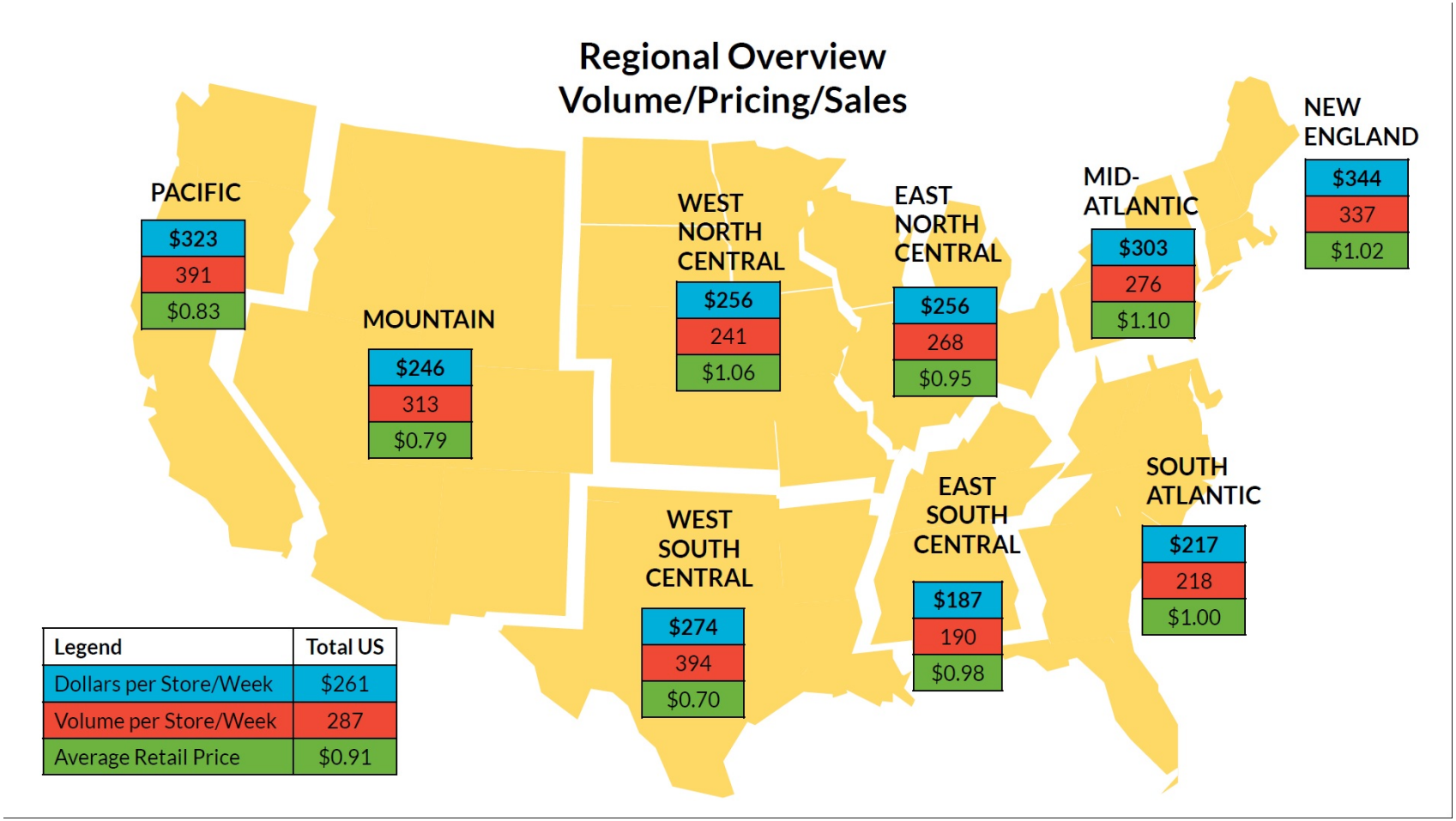
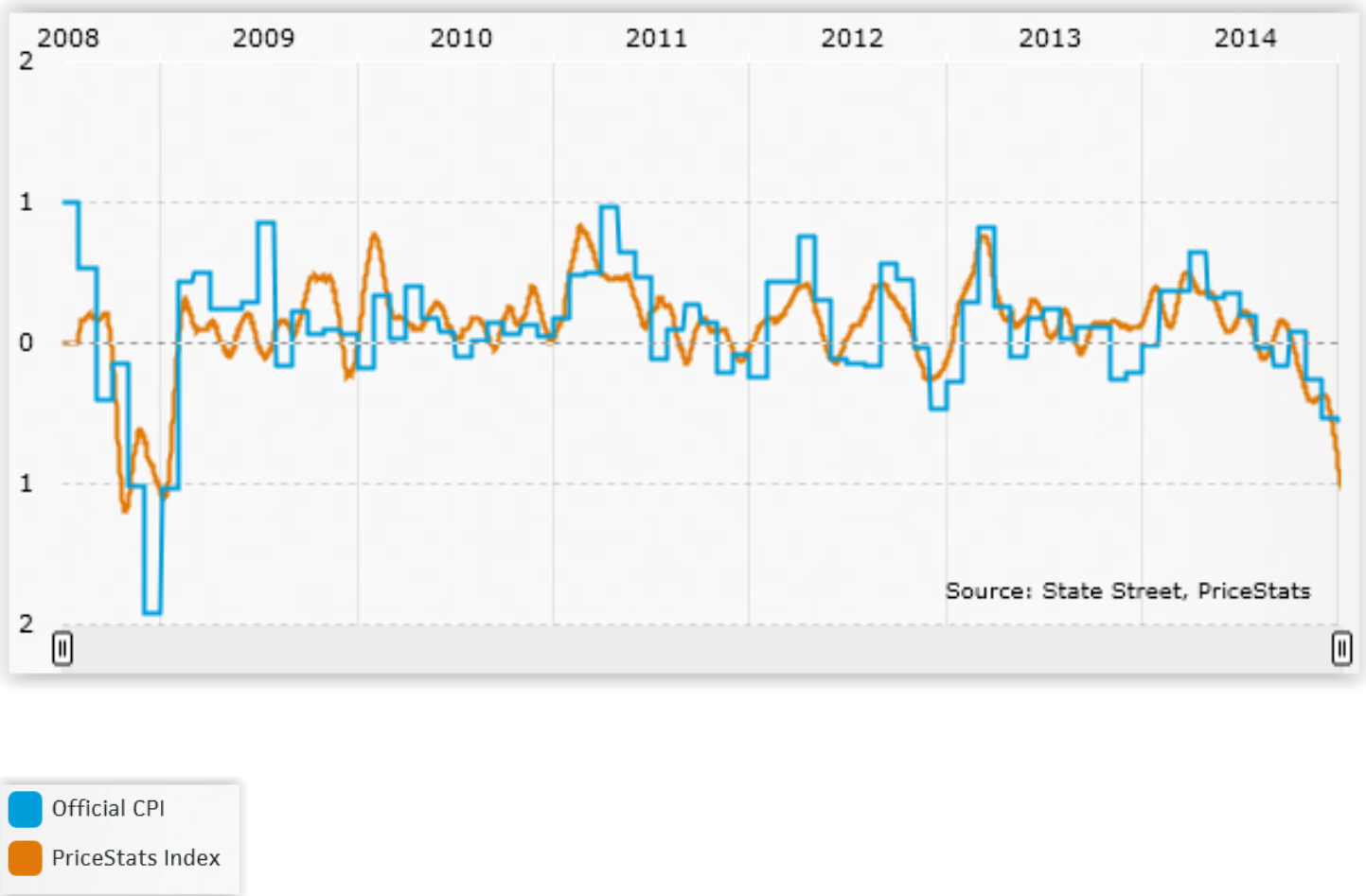
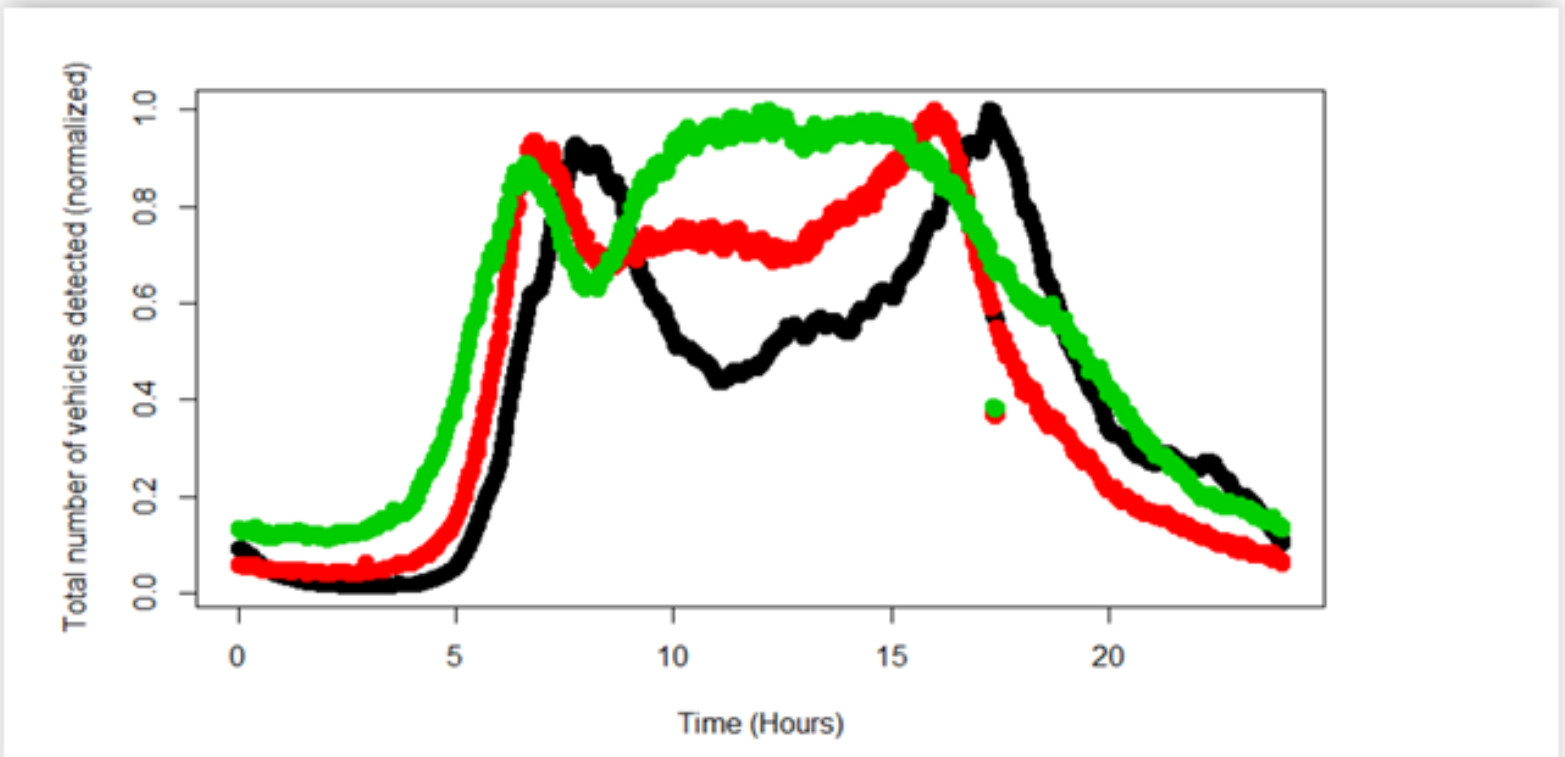


Figure: Scanner data of mango sales in grocery stores over different geographical regions; source: `https://www.mango.org/wp-content/uploads/2018/09/1st-Half-resultsEnglish.pdf`

Online Prices (AAPOR Report)



Traffic and Infrastructure (AAPOR Report)



GPS Probe Data Collection

The following figure (from FHWA, 1998) summarizes the collection of probe data

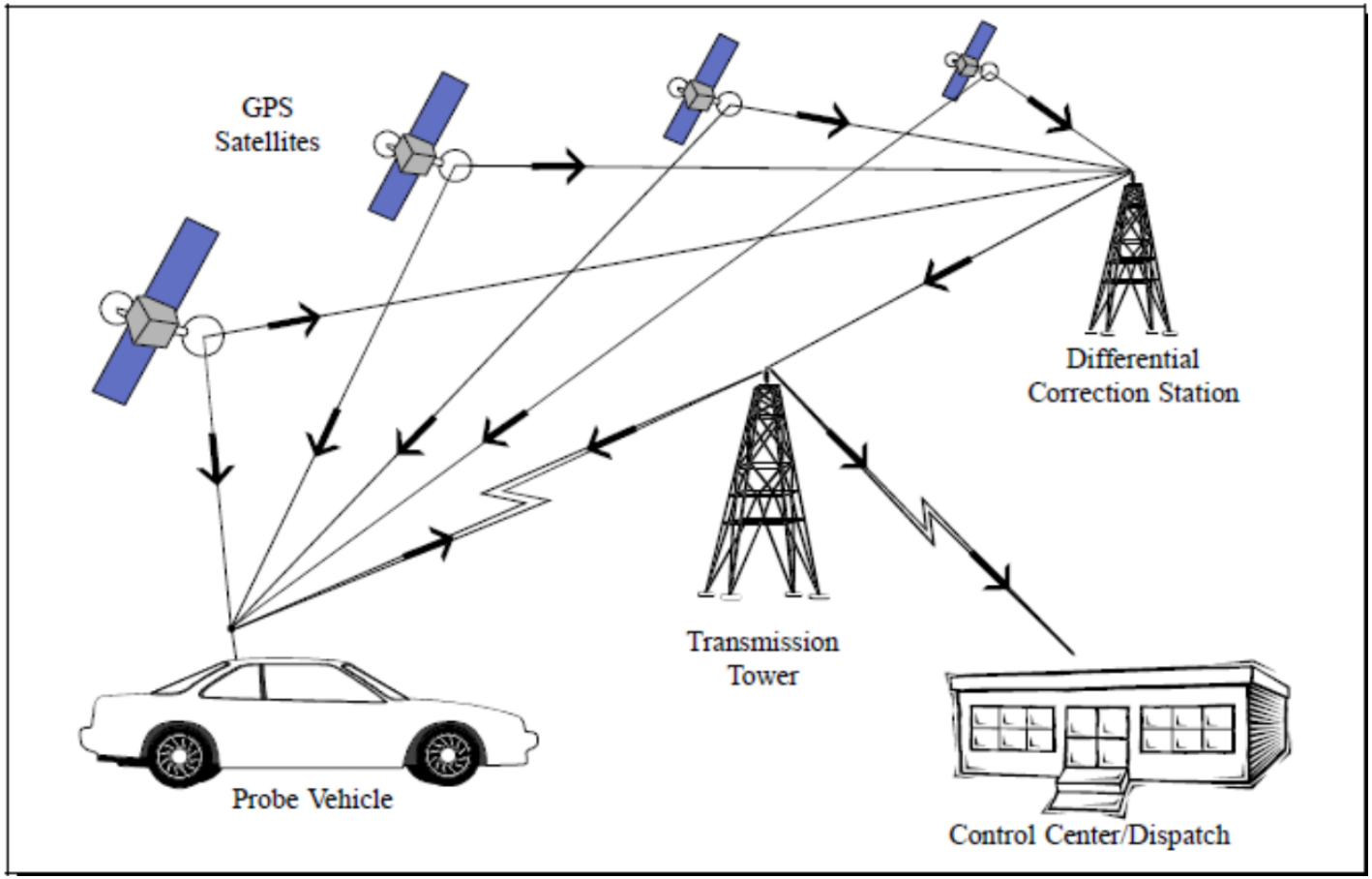
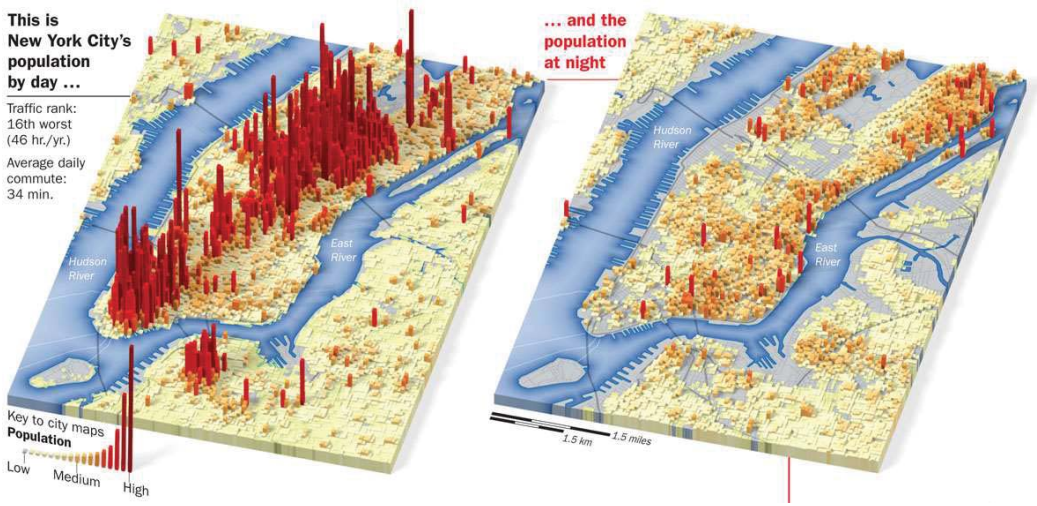


Figure: GPS Data Collection (FHWA, 1998; Source Kartika, C.S.D., 2015)

Cell Phone Data

Location data from mobile phones



Source: Pfeffermann (2017)

Small Area Modeling & Methods

Synthetic Method

Example 1: Radio Listening Survey

Ref: Hansen et al. (1953, pp 483-486)

Estimate the median number of radio stations heard during the day for over 500 counties of the USA (small areas).

Two different survey data used:

Mail Survey

- large sample (1000 families/county) from an incomplete list frame
- response rate was low (about 20%)
- estimates x_i are biased due to non-response and incomplete coverage

Personal Interview Survey

- Smaller sample size
- Sample design:
  - The 500 counties were stratified into 85 primary strata based on geographical region and the type of radio service available.
  - One county was selected from each of these 85 strata with probability proportional to the estimated number of families in the county.
  - A subsample of area segments was selected from each sampled county and the families within the selected area segments were interviewed.

- Nonresponse and coverage error properties were better than the mail survey.
- Reliable estimates y_i for the 85 sampled counties were available, but no estimate can be produced for the remaining 415 counties.
- Using (y_i, x_i) for the 85 sampled counties, the following fitted line was obtained:

$$\hat{y}_i = 0.52 + 0.74 x_i.$$

- Use y_i for the 85 sampled counties and $\hat{y}_i$ for the rest.

Figure: Comparison of median numbers of stations heard during the day as estimated from mail and interview surveys in selected primary sampling units.

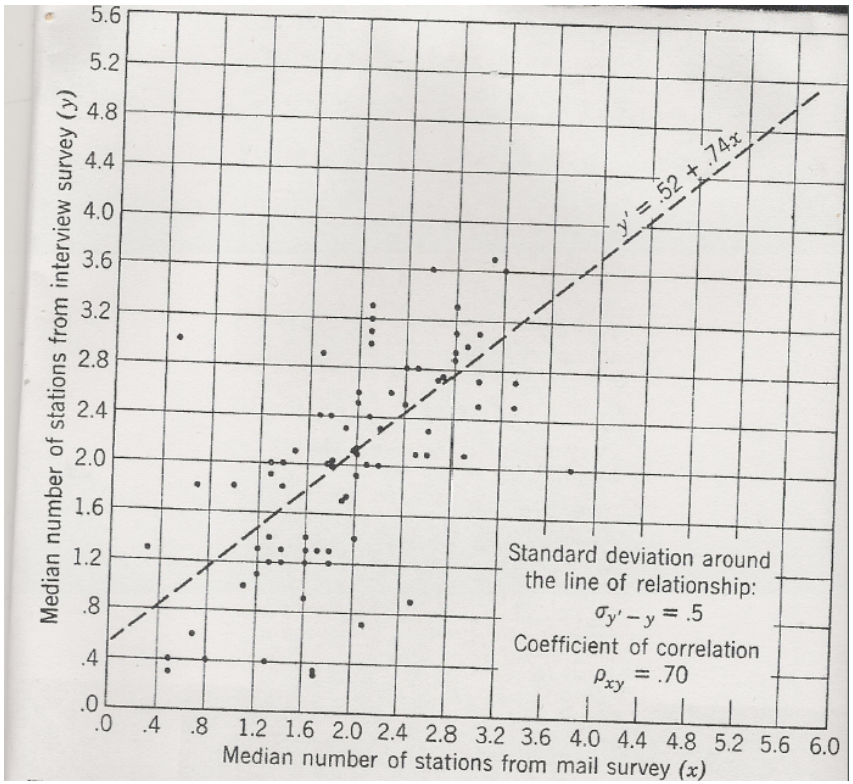


Table: Synthetic estimation using National Natality Survey

Race	Age	N_{ig}	$p_{.g}$	$N_{ig}p_{.g}$
White	Under 20	16382	0.216	3539
	20-24	44100	0.214	9437
	25-29	46421	0.222	10305
	30-34	22400	0.224	5018
	35+	5896	0.244	1439
All Other	Under 20	5493	0.173	950
	20-24	7657	0.167	1279
	25-29	5063	0.19	962
	30+	3387	0.266	901
		156799		33830

- N_{ig}: Female population size for the gth race x age-group for the ith state. We consider the state of Pennsylvania and the data are obtained from the hospital registration system.

- $p_{.g}$: national level direct estimate of the proportion of jaundiced infants whose mother is in the gth group. The data is obtained from the 1980 National Natality Survey.

- A synthetic estimate of the percentage of jaundiced infants in Pennsylvania: $p_i^s = \frac{33830}{156799} * 100 = 21.6\%$.

- Estimate of total number of jaundiced infants in Pennsylvania$=\sum_g N_{ig} p_{.g} = 33,830$.

Micro-simulation

Ref: Elbers et al. (2003)

- x_j: vector of known auxiliary variables for the jth unit of the finite population, $j \in U$.
- Sample: (y_j, x_j), $j \in s$.
- Fit the following simple linear regression model

$$y_j = x_j^T \beta + \epsilon_j, \; j \in s.$$

Let the fitted values and the residuals be:

$$\hat{y}_j = x_j^T \hat{\beta} + \epsilon_j, \; j \in s,$$

where $\hat{\beta}$ is the ordinary least squares estimator of β.

- Generate the micro-simulation census file using

$$y_j^* = \hat{y}_j + e_j^*, \; j \in U,$$

where $\{e_j^*\}$ are generated from $\{e_j, \; j \in s\}$ by SRSWR (URS).

- Repeat the process B times independently to obtain B micro-simulation census files

$$\{y^*_{bj},\ b = 1, \cdots, B;\ j \in U\}$$

- The estimator of $\bar{Y}_i$ is given by:

$$\bar{Y}^*_i = B^{-1} \sum_{b=1}^{B} \bar{Y}^*_{bi},$$

where $\bar{Y}^*_{bi} = N_i^{-1} \sum_{j \in U_i} y^*_{bj}$.

- The MSE is estimated by:

$$mse(\bar{Y}^*_i) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\bar{Y}^*_{bi} - \bar{Y}^*_i\right)^2.$$

EB and HB for an area level model

Area Level Modeling

For $i = 1, \cdots, m$, consider the following Bayesian model

$$\textit{Level 1: (Sampling Distribution)}: \ y_i|\theta_i \sim N(\theta_i, \psi_i);$$
$$\textit{Level 2: (Prior Distribution)}: \ \theta_i \sim N(x_i'\beta, A)$$

where

- m : number of small area;
- y_i : direct survey estimate of θ_i;
- θ_i : true mean for area i;
- x_i : $p \times 1$ vector of known auxiliary variables;
- ψ_i: known sampling variance of the direct estimate;
- The $p \times 1$ vector of regression coefficients β and model variance A are unknown.

Bayes Estimator of θ_i

Inferences based on the posterior distribution of θ_i:

$$\theta_i | y; \beta, A \overset{ind}{\sim} N(\hat{\theta}_i^B,\ \sigma_i^2(A)),$$

where

- $\hat{\theta}_i^B = (1 - B_i) y_i + B_i x_i' \beta$
- $B_i = \frac{\psi_i}{A + \psi_i}$
- $\sigma_i^2(A) = (1 - B_i)\psi_i$

Emiprical Bayes (EB) Estimator of θ_i

- Treat the hyperparameters β and A fixed and estimate them by consistent estimators (e.g., ANOVA, ML, REML, adjusted ML).
- Substituting the hyperparameters by their respective estimators in the Bayes estimator of θ_i, one obtains an empirical Bayes (EB) estimator of θ_i.
- An empirical Bayes estimator of θ_i can be also motivated from the classical prediction approach in a linear mixed model. Note that we can also write the Bayesian model as:

$$y_i = \theta_i + e_i = x_i'\beta + v_i + e_i,\ i = 1, \cdots, m,$$

where the random effects $\{v_i,\ i = 1, \cdots, m\}$ and the errors $\{e_i,\ i = 1, \cdots, m\}$ are independent with $v_i \sim N(0, A)$, and $e_i \sim N(0, \psi_i)$.
The best predictor of θ_i is identical to the Bayes estimator of θ_i and hence an empirical Bayes estimator of θ_i can be also called an empirical best (EB) predictor of θ_i.

Hierarchical Bayes (HB) Estimator of θ_i

- Put priors, possibly non-informative flat priors, on the hyperparameters β and A.
- The inference is based on the posterior distribution of the target parameter.
- Point estimate of θ_i: $E(\theta_i|y) = E\left[E(\theta_i|y;\beta,A)|y\right]$
- Measure of uncertainty of the point estimate:

$$V(\theta_i|y) = E\left[V(\theta_i|y;\beta,A)|y\right] + V\left[E(\theta_i|y;\beta,A)|y\right].$$

- The credible interval of θ_i is obtained using the posterior distribution of θ_i.
- For this simple model, the HB method can be implemented using on-dimensional numerical integration. For more complex model, The Monte Carlo Markov Chain is used.

Example 1: Estimation of per-capita income of small places

Ref: Fay and Herriot (1979, JASA); Rao and Molina (2015)

- The US Census Bureau provides the US Treasury department with estimates of per-capita income (PCI) and other statistics for states and other local governments receiving funds under the general revenue sharing program.

- The Treasury department uses these statistics to determine allocations to the local government units.

Example 1 Continued

- Initial Method: Current PCI estimate for a place = (1970 census estimate of PCI in 1969 based on about 20% sample) x (the ratio of an administrative estimate of PCI in the current year and a similarly derived estimate for 1969.)

- Coefficient of variation (CV): about 13% for a place of 500 persons and about 30% for a place of 100 persons.

Example 1 Continued

- Synthetic method: substitutes the corresponding county estimates in their place.
- This solution is unsatisfactory (why?)

Example 1 Continued

- The EBLUP used by Fay and Herriot (1979): a weighted average of the direct estimator and a regression synthetic estimator.
- Fay and Herriot demonstrated that the EBLUP estimates of log(PCI) for small places have average error smaller than the census estimates or the county estimates.
- The Fay-Herriot EBLUP method was adopted by the Census Bureau in 1974 to form updated PCI estimates for small places.

Example 1 Continued

Sampling variances:

- CV of the direct estimate of PCI $\approx 3/\sqrt{\hat{N}_i}$, where $\hat{N}_i$ is the weighted sample count.
- Use logarithmic transformation of direct estimate with sampling variance estimated as $\psi_i = 9/\hat{N}_i$

Auxiliary variables considered:

(1) x_2=PCI for the county

(2) x_3=value of owner-occupied housing for the place

(3) x_4=value of owner-occupied housing for the county

(4) x_5= IRS-adjusted gross income per exemption from the 1969 tax returns for the place

(5) x_6= IRS-adjusted gross income per exemption from the 1969 tax returns for the county

Example 1 Continued

Models Considered

- Model (1): $p = 2$ with $x_1 = 1$ (corresponding to the intercept), x_2
- Model (2): $p = 4$ with x_1, x_2, x_3 and x_4.
- Model (3): $p = 4$ with x_1, x_2, x_5, and x_6
- Model (4): $p = 6$ with x_1-x_6

Estimated Model Variances for Different Models

State	Model (1)	Model (2)	Model (3)	Model (4)
Illinois	0.036	0.032	0.019	0.017
Iowa	0.029	0.011	0.017	0.000
Kansas	0.064	0.048	0.016	0.020
Minnesota	0.063	0.055	0.014	0.019
Missouri	0.061	0.033	0.034	0.017
Nebraska	0.065	0.041	0.019	0.000
North Dakota	0.072	0.081	0.020	0.004
South Dakota	0.138	0.138	0.014	*
Wisconsin	0.042	0.025	0.025	0.004

*Not fitted because some auxiliary variables for Model (4) were unavailable for several places in South Dakota

Source: Adapted from Table 1 of Fay and Herriot (1979) and Table 6.1 of Rao and Molina (2015)

Example 1 Continued: External Evaluation

Evaluation:

The Census Bureau conducted complete censuses of a random sample of places and townships in 1973 and collected income data for 1972 on a 100% basis.

$\#$ of places with population size < 500 : 17.

$\#$ of places with population size between 500 and 999: 7.

Estimates for 1972 were obtained by multiplying the estimates by updating factors f_i

Average Percent Difference

N	$\hat{\bar{Y}}_i$	$\hat{\bar{Y}}_i^*$	$\hat{\bar{Y}}_i^C$
< 500	28.6	22.0	31.6
500-999	19.1	15.6	19.3

where $\hat{\bar{Y}}_i^C =$ County estimate.

Example 1 Continued: External Evaluation

Percent Difference

Special Census Area	Direct Est.	EBLUP	County Est.
1	10.2	14.0	12.9
2	4.4	10.3	30.9
3	34.1	26.2	9.1
4	1.3	8.3	24.6
5	34.7	21.8	6.6
6	22.1	19.8	14.6
7	14.1	4.1	18.7
8	18.1	4.7	25.9
...	...	...	...
17	51.4	14.4	23.7
Average	28.6	22.0	31.6

Source: The table is adapted from Table 3 of Fay and Herriot (1979) and Table 6.2 of Rao and Molina (2015)

Example 2: SAIPE Data Analysis

Ref: Bell and Franco (2017)
`https://www.census.gov/srd/csrmreports/byyear.html`.
Erciulescu, Franco and Lahiri (2018)

- **cps93** – The direct CPS estimated poverty rates for related children ages 5-17.
- **irspr93** – The pseudo-poverty rates tabulated from IRS tax data. These are defined as the number of child tax exemptions for poor households divided by the total number of child tax exemptions.
- **irsnf93** – The tax non-filer rates tabulated from IRS tax data, defined as the difference between the estimated population and number of tax exemptions under age 65, divided by the estimated population under age 65.
- **fs93** – The Food Stamp participation proportions. This variable is the average monthly number of individuals receiving food stamps over a 12-month period, as a percentage of the population.

- **smpsize** – The CPS sample size (number of interviewed households).
- **fnlse** – The GVF estimates of sampling standard errors from the CPS. These are computed using the GVF developed by Otto and Bell (1995), using an iterative procedure that alternates between estimation of model parameters via maximum likelihood and estimation of the sampling standard errors.
- **cen89rsd** – The residuals obtained by fitting a Fay-Herriot model to the estimates of children in poverty from the 1990 census, with analogous covariates to those used here but for the year 1989.

In the area level model, for state i

- y_i: **cps93**
- x_i: an intercept term, and **irspr93**, **irsnf93**, **fs93**, **cenrsd**
- ψ_i: **fnlse**^{2}.

The analysis here is done with the "sae" package in R (Molina and Marhueda, 2015)

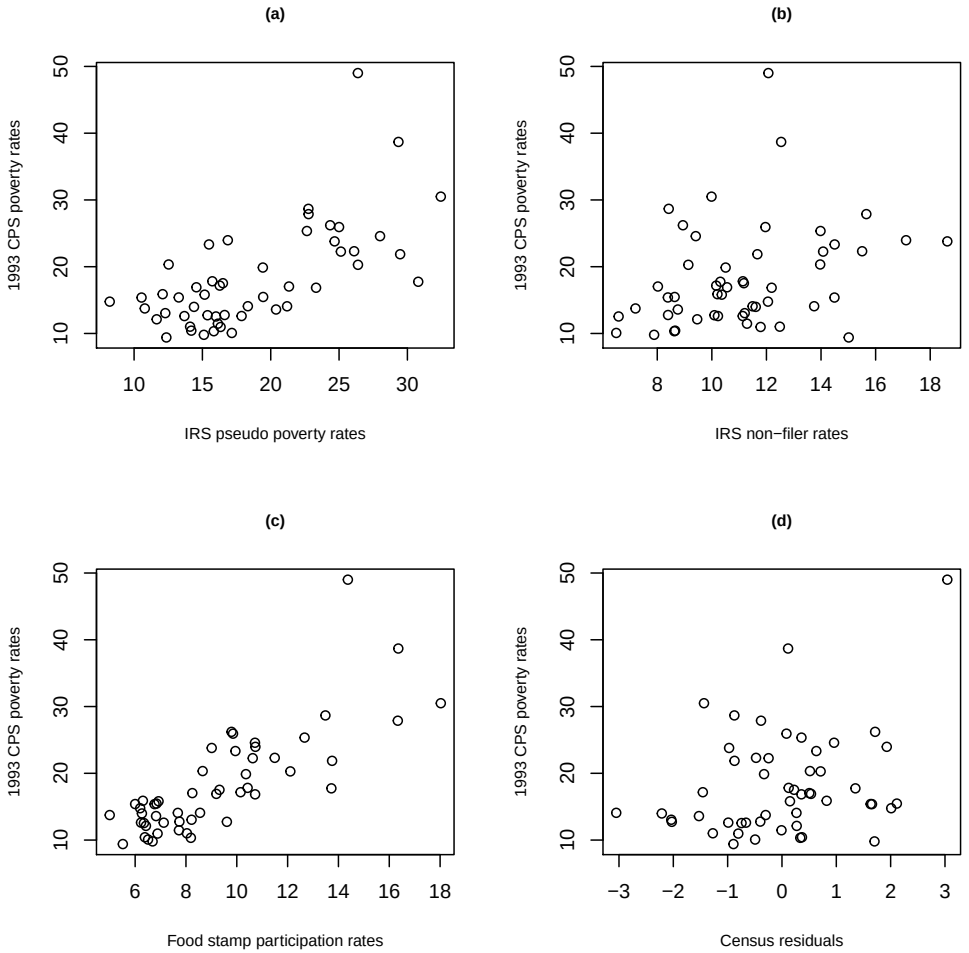


Figure: CPS poverty rates for school-aged children plotted against (a) IRS pseudo-poverty rates, (b) IRS non-filer rates (c) Food Stamp (SNAP) participation rates, (d) Census 1990 residuals

Variable	Coefficient	S.E.	t	$Pr > \|t\|$
Intercept	−3.477	2.224	−1.564	0.118
IRS pseudo-poverty rate	0.267	0.125	2.144	0.032
IRS non-filer rate	0.509	0.156	3.261	0.001
Food stamp participation rate	1.185	0.268	4.429	<0.001
Census residuals	1.261	0.413	3.050	0.002

Model	Regressors	Model Variance	AIC
M1	IRS pseudo poverty rate	12.803	316.638
M2	IRS non-filer rate	25.307	341.525
M3	Food Stamp participation rate	6.049	294.007
M4	Census residuals	30.631	345.437
M12	IRS pseudo poverty rate, IRS non-filer rate	7.972	308.810
M13	IRS pseudo poverty rate, Food Stamp participation rate	6.051	295.332
M14	IRS pseudo poverty rate, Census residuals	9.741	311.074
M23	IRS non-filer rate, Food Stamp participation rate	3.449	290.033
M24	IRS non-filer rate, Census residuals	24.310	339.345
M34	Food Stamp participation rate, Census residuals	5.468	289.568
M123	IRS pseudo poverty rate, IRS non-filer rate, Food Stamp participation rate	3.061	290.188
M124	IRS pseudo poverty rate, IRS non-filer rate, Census residuals	4.418	300.373
M134	IRS pseudo poverty rate, Food Stamp participation rate, Census residuals	4.85	289.749
M234	IRS non-filer rate, Food Stamp participation rate, Census residuals	3.257	285.264
M1234	IRS pseudo poverty rate, IRS non-filer rate, Food Stamp participation rate Census residuals	1.703	282.601

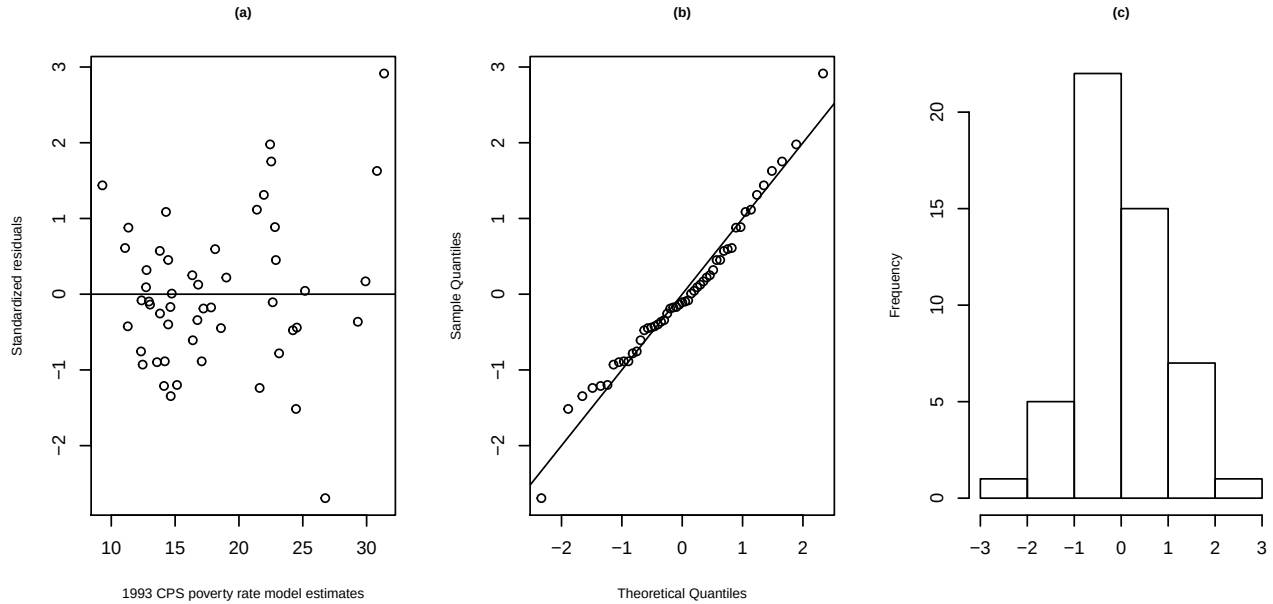


Figure: Model diagnostic plots: (a) Standardized residuals plotted against model predictions (b) Quantile to quantile plot of standardized residuals (c) Histogram of standardized residuals

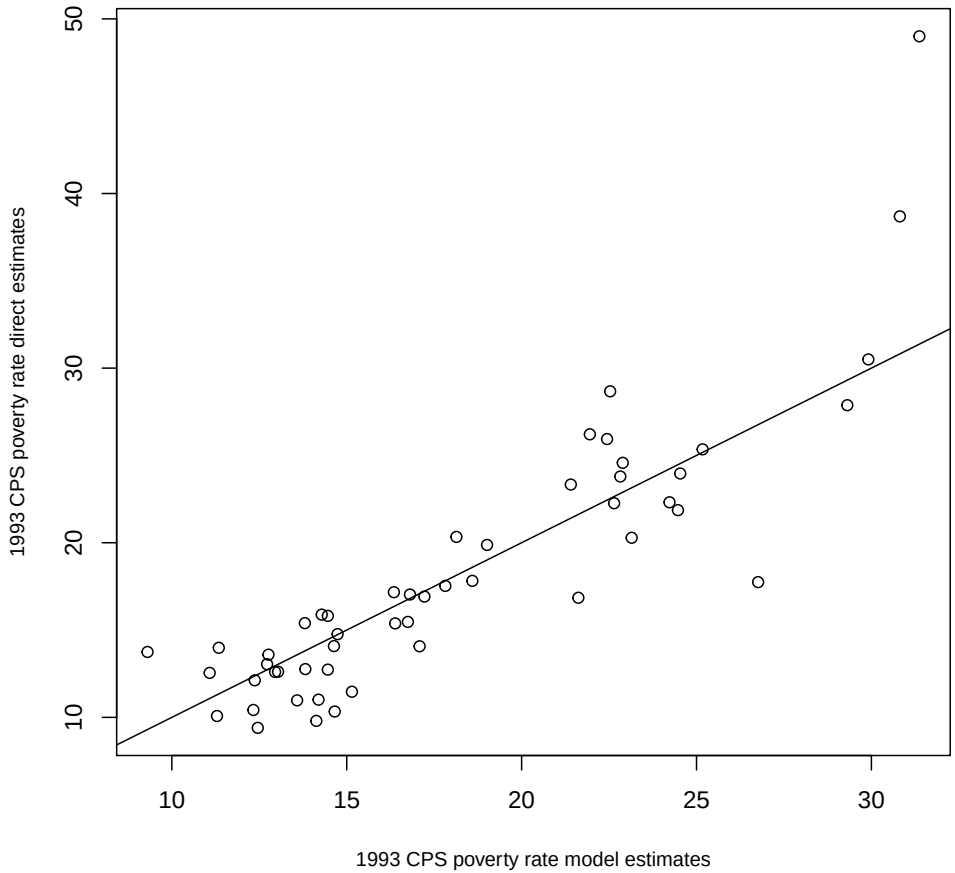


Figure: Model predictors vs. direct estimates for 1993 school-aged children in poverty based on CPS data, and $y = x$ line

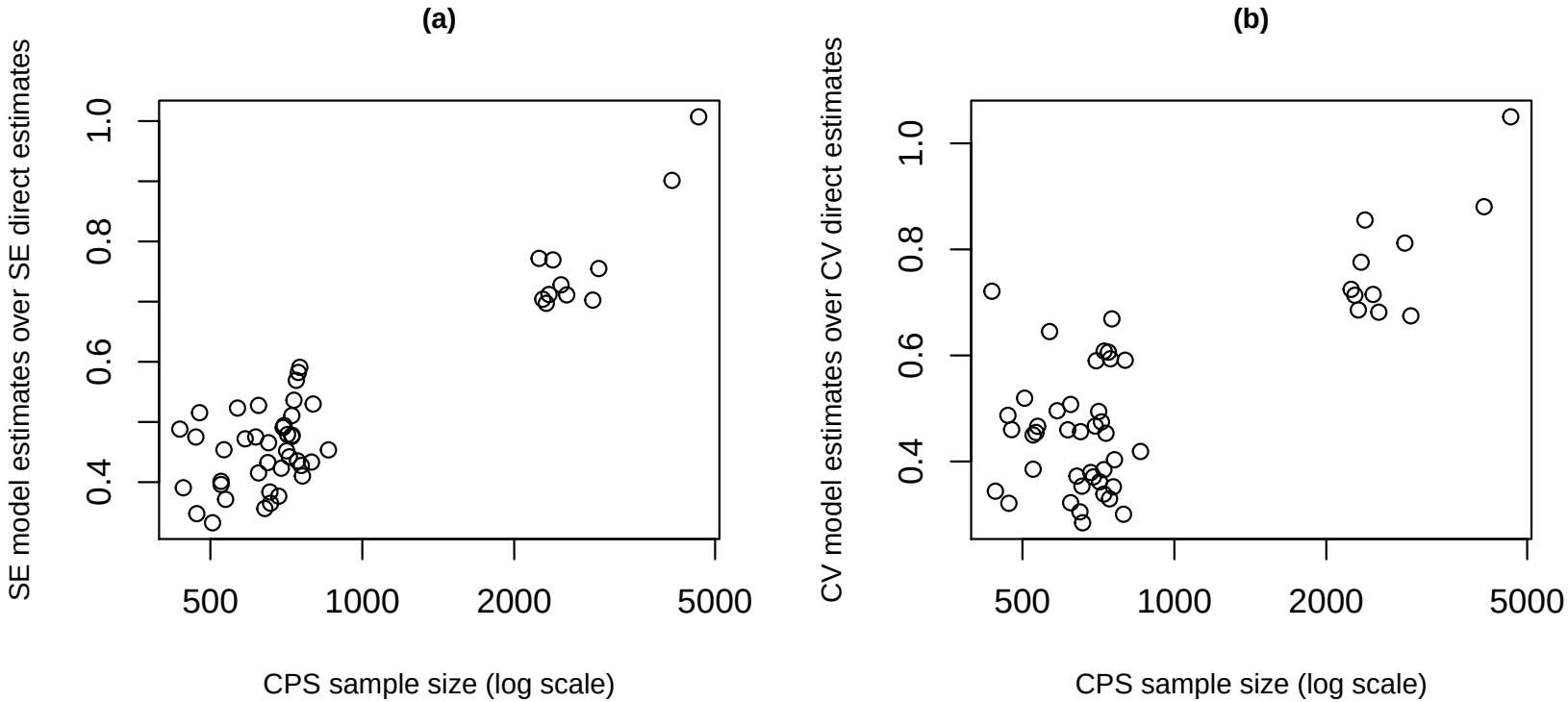


Figure: Ratios of standard errors (a) and of coefficients of variation (b) of modeled estimates over direct CPS estimates

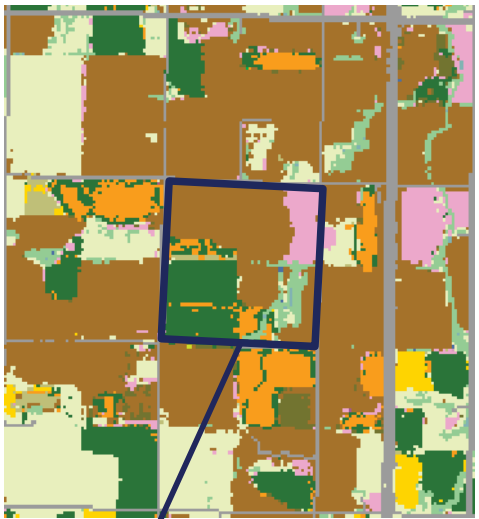
EB and HB for an unit level model

PAGE 2 **SECTION D - CROPS AND LAND USE ON TRACT** 17

How many acres are inside this blue tract boundary drawn on the photo (map)? .

Now I would like to ask about each field inside this blue tract boundary and its use during 2000.

FIELD NUMBER		01	02	03	04	05
1. Total acres in field		828	828	828	828	828
2. Crop or land use. [Specify]						
3. Occupied farmstead or dwelling		843				
4. Waste, unoccupied dwellings, buildings and structures, roads, ditches, etc.						
5. Woodland		831	831	831	831	831
6. Pasture	Permanent (not in crop rotation)	842	842	842	842	842
	Cropland (used only for pasture)	856	856	856	856	856
8. Idle cropland - Idle all during 2000		857	857	857	857	857
9. Two crops planted in this field or two uses of the same crop.		☐Yes ☐No	☐Yes ☐No	☐Yes ☐No	☐Yes ☐No	☐Yes ☐No
	[Specify second crop or use]					
	Acres	844	844	844	844	844
10. Acres left to be planted		610	610	610	610	610
11. Acres irrigated and to be irrigated. [If double cropped, include acreage of each crop irrigated.]		620	620	620	620	620
16. Winter Wheat (include cover crop)	Planted	540	540	540	540	540
17.	For grain or seed	541	541	541	541	541
18. Rye (include cover crop) [Exclude ryegrass]	Planted	547	547	547	547	547
19.	For grain or seed	548	548	548	548	548



REGRESSION VARIABLES:

	Dependent Y	Independent X
	Enumerated JAS Segments	CDL Classified Acres
Soybeans	227	273
Wheat	337	541

28

Zakzeski, A., National Agricultural Statistics Service

An Example

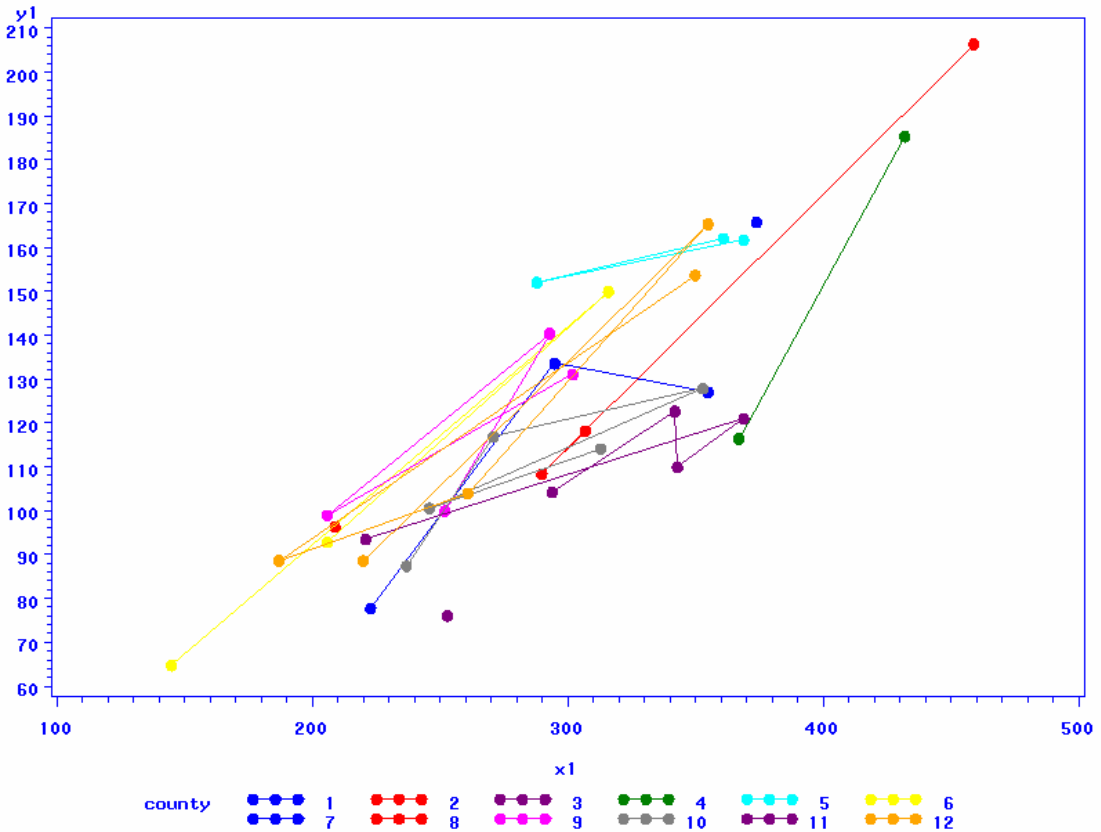
Ref: Battese, Harter and Fuller (1988 JASA)

Table 1. Survey and Satellite Data for Corn and Soybeans in 12 Iowa Counties

	No. of segments		Reported hectares		No. of pixels in sample segments		Mean number of pixels per segment*	
County	Sample	County	Corn	Soybeans	Corn	Soybeans	Corn	Soybeans
Cerro Gordo	1	545	165.76	8.09	374	55	295.29	169.70
Hamilton	1	566	96.32	106.03	209	218	300.40	196.65
Worth	1	394	76.08	103.60	253	250	289.60	205.28
Humboldt	2	424	185.35	6.47	432	96	290.74	220.22
			116.43	63.82	367	178		
Franklin	3	564	162.08	43.50	361	137	318.21	188.06
			152.04	71.43	288	206		
			161.75	42.49	369	165		
Pocahontas	3	570	92.88	105.26	206	218	257.17	247.13
			149.94	76.49	316	221		
			64.75	174.34	145	338		
Winnebago	3	402	127.07	95.67	355	128	291.77	185.37
			133.55	76.57	295	147		
			77.70	93.48	223	204		
Wright	3	567	206.39	37.84	459	77	301.26	221.36
			108.33	131.12	290	217		
			118.17	124.44	307	258		
Webster	4	687	99.96	144.15	252	303	262.17	247.09
			140.43	103.60	293	221		
			98.95	88.59	206	222		
			131.04	115.58	302	274		
Hancock	5	569	114.12	99.15	313	190	314.28	198.66
			100.60	124.56	246	270		
			127.88	110.88	353	172		
			116.90	109.14	271	228		
			87.41	143.66	237	297		
Kossuth	5	965	93.48	91.05	221	167	298.65	204.61
			121.00	132.33	369	191		
			109.91	143.14	343	249		
			122.66	104.13	342	182		
			104.21	118.57	294	179		
Hardin	6	556	88.59	102.59	220	262	325.99	177.05
			88.59	29.46	340	87		
			165.35	69.28	355	160		
			104.00	99.15	261	221		
			88.63	143.66	187	345		
			153.70	94.49	350	190		

* The mean number of pixels of a given crop per segment in a county is the total number of pixels classified as that crop, divided by the number of segments in that county.

Fig 2: Plot of Corn Hectares versus Corn Pixels by County



This plot also reflects the strong relationship between the reported hectares of corn and the number of pixels of corn for counties separately. But the slopes and/or intercepts seem differ by county.

Estimation of average hectares of corn for counties

- Each county was divided into area segments. A segment is a primary sampling unit. Segments are about 250 hectares.
- y_{ij}: the number of hectares of corn in the jth segment of the ith county as reported by the farm operators in the June Enumerative Survey.
- $x'_{ij} = (1, x_{1ij}, x_{2ij})$, where x_{1ij} (x_{2ij}) is the number of *pixels* ("picture elements") classified as corn (soybean) in the jth segment of the ith county. A pixel is about 0.45 hectares. This auxiliary data is available not only for the sampled segments but also for all segments in each counties.
- $\bar{X}' = (1, \bar{X}_{1i}, \bar{X}_{2i})$, where $\bar{X}_{1i}$ ($\bar{X}_{2i}$) is the mean number of pixels per segment classified as corn (soybeans) for county i. This is the total number of pixels classified as corn divided by the number of pixels in that county.

Estimation of average hectares of corn for counties

- BHF considered estimation of average hectares of corn and soybeans for 12 counties in North-Central based on the 1978 June Enumerative Survey and satellite data.
- y_{ij} : hectarage of corn for the jth segment of the i county $(i = 1, \cdots, m;\ \ j = 1, \cdots, N_i)$
- We are interested in estimating the average hectares of corn for county i $(i = 1, \cdots, m)$:

$$\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}.$$

BHF Fixed effects Models analyses

We begin by exploring the fixed effects model. We fit 3 different models:

- *Model 1:* $y_i = \beta_0 + \beta_1 * corn + e_i$
- *Model 2:* $y_i = \beta_0 + \beta_2 * soybean + e_i$
- *Model 3:* $y_i = \beta_0 + \beta_1 * corn + \beta_2 * soybean + e_i$

We ran Step-wise regression on the full model (Model 3) based on BIC (correction factor, $k = \log(n = 37)$ in AIC) and we got the following result:

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
y ~ x1 + x2

Final Model:
y ~ x1


  Step Df Deviance Resid. Df Resid. Dev      AIC
1                         34   12106.62 225.0846
2 - x2  1 55.41188        35   12162.03 221.6426
```

Figure: the BIC criterion chooses Model 1

```
Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-47.991 -13.662   2.316  13.657  35.305

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.81871   13.48779   0.506    0.616
x1           0.38165    0.04417   8.641 3.34e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.64 on 35 degrees of freedom
Multiple R-squared:  0.6809,    Adjusted R-squared:  0.6718
F-statistic: 74.67 on 1 and 35 DF,  p-value: 3.344e-10
```

Figure: summary of `lm` for Model 1

.

```
Call:
lm(formula = y ~ x2)

Residuals:
    Min      1Q  Median      3Q     Max
-67.678 -13.961  -3.132  21.262  47.032

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 183.14999   13.40218  13.666 1.35e-15 ***
x2           -0.30899    0.06265  -4.932 1.97e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.35 on 35 degrees of freedom
Multiple R-squared:   0.41,     Adjusted R-squared:  0.3932
F-statistic: 24.33 on 1 and 35 DF,  p-value: 1.969e-05
```

Figure: summary of `lm` for Model 2

.

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-50.361 -13.431   0.721  13.349  35.194

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.29100   32.12735   0.569    0.573
x1           0.36194    0.06705   5.398 5.22e-06 ***
x2          -0.02759    0.06995  -0.394    0.696
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.87 on 34 degrees of freedom
Multiple R-squared:  0.6823,    Adjusted R-squared:  0.6636
F-statistic: 36.51 on 2 and 34 DF,  p-value: 3.417e-09
```

Figure: summary of `lm` for Model 3

.

Clearly we see that in terms of predictive power (R^2_{adj}) as well as BIC, Model 1 is superior to the other fixed effects models 2 and 3.

Nested Error Regression Model

For $i = 1, \cdots, m; \;\; j = 1, \cdots, N_i$

$$y_{ij} = x_{ij}'\beta + v_i + e_{ij},$$

where x_{ij} is a $p \times 1$ column vector of known auxiliary variables;β is a $p \times 1$ column vector of unknown regression coefficients; $\{v_i\}$ and $\{e_{ij}\}$ are all independent with $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ and $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$

Estimation of corn hectares: EBLUP

- Battese et al. (1988) fitted the following fitted nested error model for corn:

$$\hat{y}_{ij} = 51 + 0.329x_{1ij} - 0.134x_{2ij},\ \hat{\sigma}_e^2 = 150,\ \hat{\sigma}_v^2 = 140$$

- All regression coefficients are significant; The among-county variance σ_v^2 is significant at the 10% level.
- The Sapiro-Wilks test and normal probability plot applied on transformed residuals fail to reject the normality assumption of the assumed nested error model.
- An approximate F-test reveals that slopes are not significantly different for within and among counties.

BHF Mixed effects Models analyses

Next we move on to explore the following 3 Mixed-Effects Models:

- *Model 4:* $y_i = \beta_0 + \beta_1 * corn + V_i + e_i$
- *Model 5:* $y_i = \beta_0 + \beta_2 * soybean + V_i + e_i$
- *Model 6:* $y_i = \beta_0 + \beta_1 * corn + \beta_2 * soybean + V_i + e_i$

where, $V_i \sim N(0, \sigma_v^2)$ independently of $e_i \sim N(0, \sigma_e^2)$.
We use Stan interfaced with R to fit the above models with flat priors on β's and Cauchy prior on σ_v.

We use `loo` (Leave-One-Out CV and ELPD criterion) to choose the best model (Vehtari et al.).

	elpd_diff	se_diff	elpd_loo	p_loo	looic
loo4.new	0.0	0.0	-177.8	2.8	355.7
loo6.new	-0.6	0.9	-178.5	3.6	356.9
loo5.new	-3.6	3.2	-181.5	4.5	363.0

Figure: `comparison of the different models using loo`

.

Model 4 is selected as the best model; closely contested by Model 6 (only 0.6 difference in elpd from Model 4). Clearly Model 5 is the worst, having the maximum difference (3.6) in elpd from Model 4.

Parameter	mean	sd	se_mean	2.5%	50%	97.5%
beta0	0.83	9.92	0.08	-18.54	0.77	20.26
beta1	0.40	0.04	0.00	0.33	0.40	0.48
V1	1.39	6.17	0.06	-7.90	0.30	17.94
V7	-0.50	5.05	0.03	-12.60	-0.15	9.35
V11	-1.25	5.15	0.04	-15.12	-0.36	7.04
V12	-1.38	5.28	0.05	-15.61	-0.40	7.07

Table: Parameter estimates, posterior standard error, Monte-Carlo standard error and confidence limits for Model 4 (intercept, corn and random effects V_i)

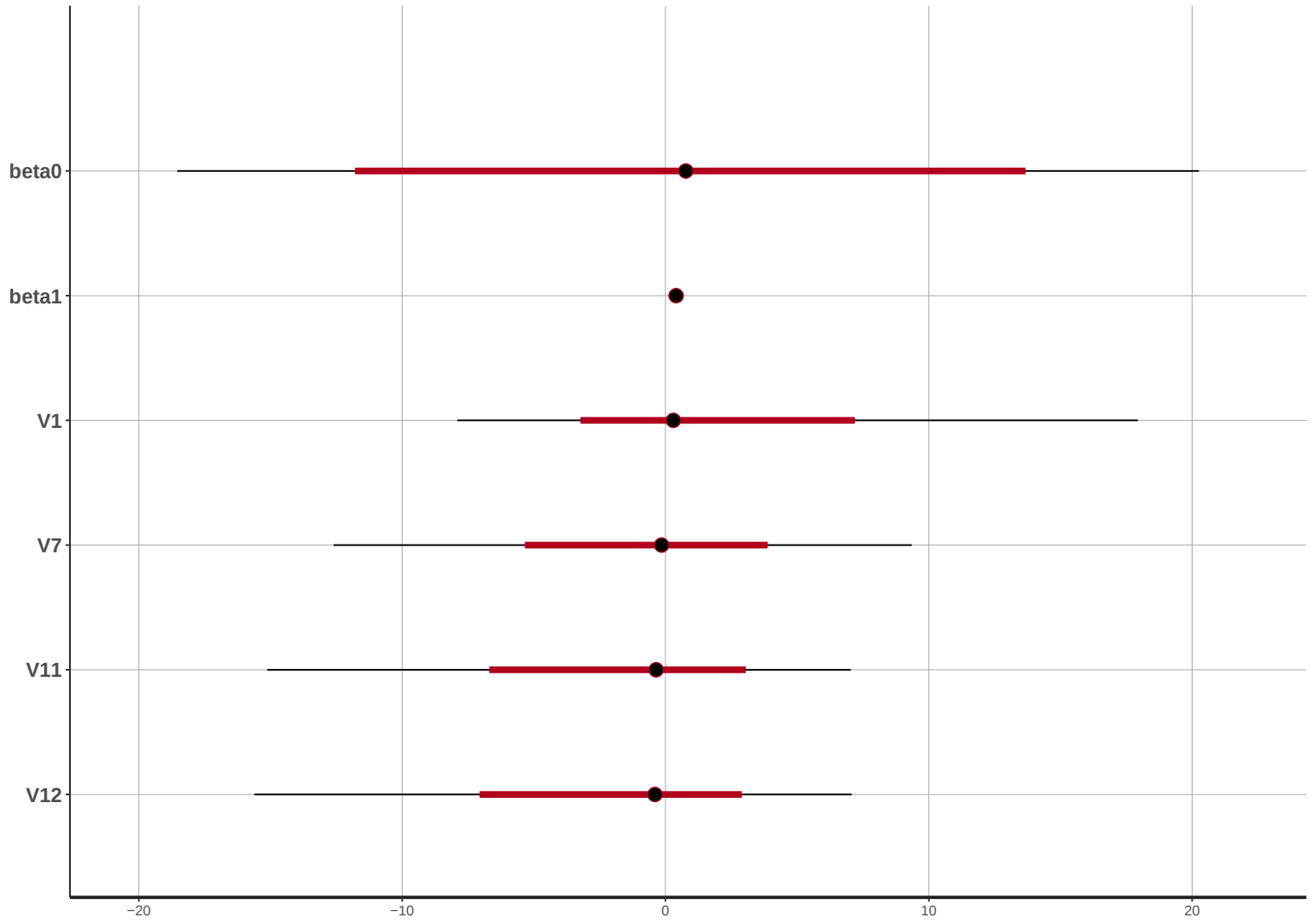


Figure: CI plot for parameter estimates from Model 4.

Parameter	mean	sd	se_mean	2.5%	50%	97.5%
theta[1]	121.60	8.37	0.07	107.38	120.98	140.85
theta[2]	121.42	8.01	0.04	104.49	121.67	136.34
theta[3]	116.55	8.06	0.06	98.76	117.02	130.75
theta[4]	120.10	8.18	0.08	106.33	119.42	139.56
theta[5]	131.62	8.50	0.09	117.27	130.92	151.51
theta[6]	104.59	6.80	0.03	91.07	104.59	118.32
theta[7]	118.28	7.12	0.03	103.89	118.36	132.21
theta[8]	124.15	7.82	0.07	110.37	123.65	142.08
theta[9]	107.77	6.90	0.05	95.25	107.40	122.91
theta[10]	125.85	7.58	0.07	109.26	126.35	139.48
theta[11]	120.31	7.06	0.04	105.28	120.63	133.43
theta[12]	131.24	7.48	0.05	115.56	131.50	145.20

Table: Estimates of θ_i's – the estimates of small area means – from Model 4

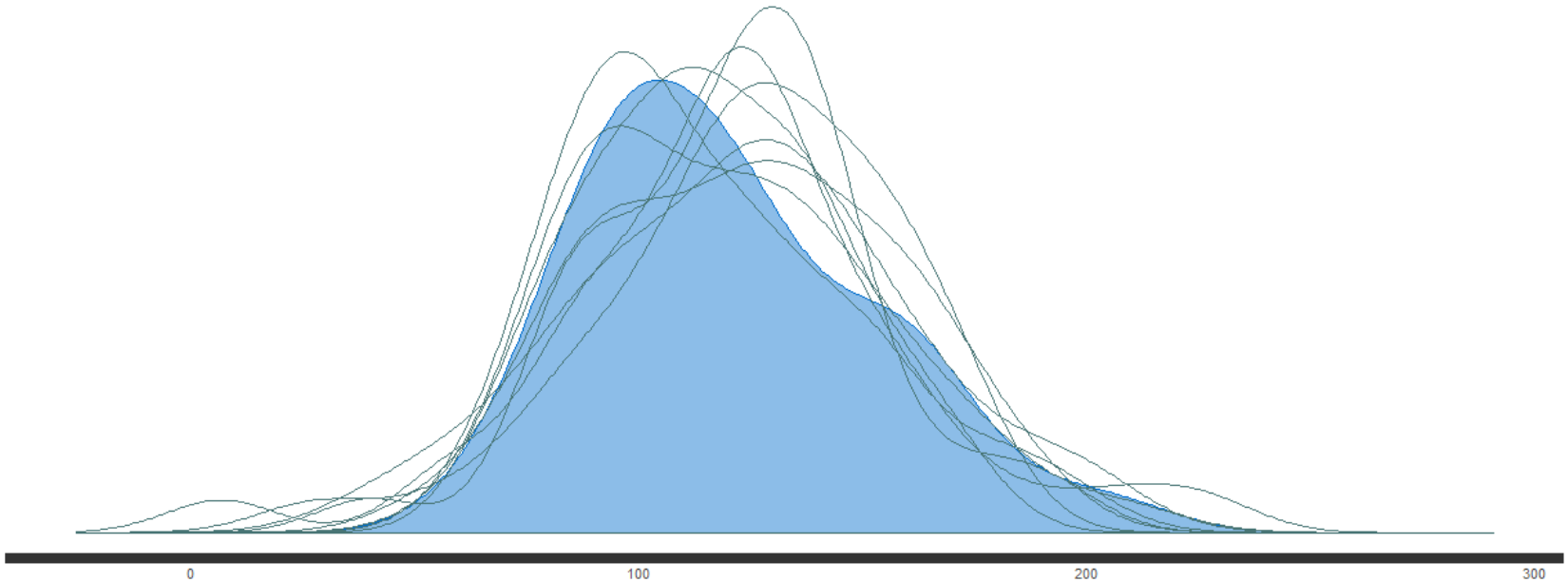


Figure: Posterior predictive simulations from Model 4. The shaded density is of the actual data.

Parameter	mean	sd	se_mean	2.5%	50%	97.5%
beta0	0.09	10.10	0.30	-19.64	0.05	19.15
beta1	0.32	0.11	0.00	0.11	0.32	0.54
beta2	0.13	0.16	0.00	-0.19	0.13	0.43
V1	1.50	6.54	0.06	-7.85	0.25	19.14
V7	-0.34	5.17	0.03	-12.60	-0.02	9.87
V11	-1.24	5.16	0.05	-15.30	-0.27	7.07
V12	-0.83	5.39	0.04	-14.60	-0.14	8.72

Table: Parameter estimates, posterior standard error, Monte-Carlo standard error and confidence limits for Model 6 (intercept, corn, soybean and random effects V_i)

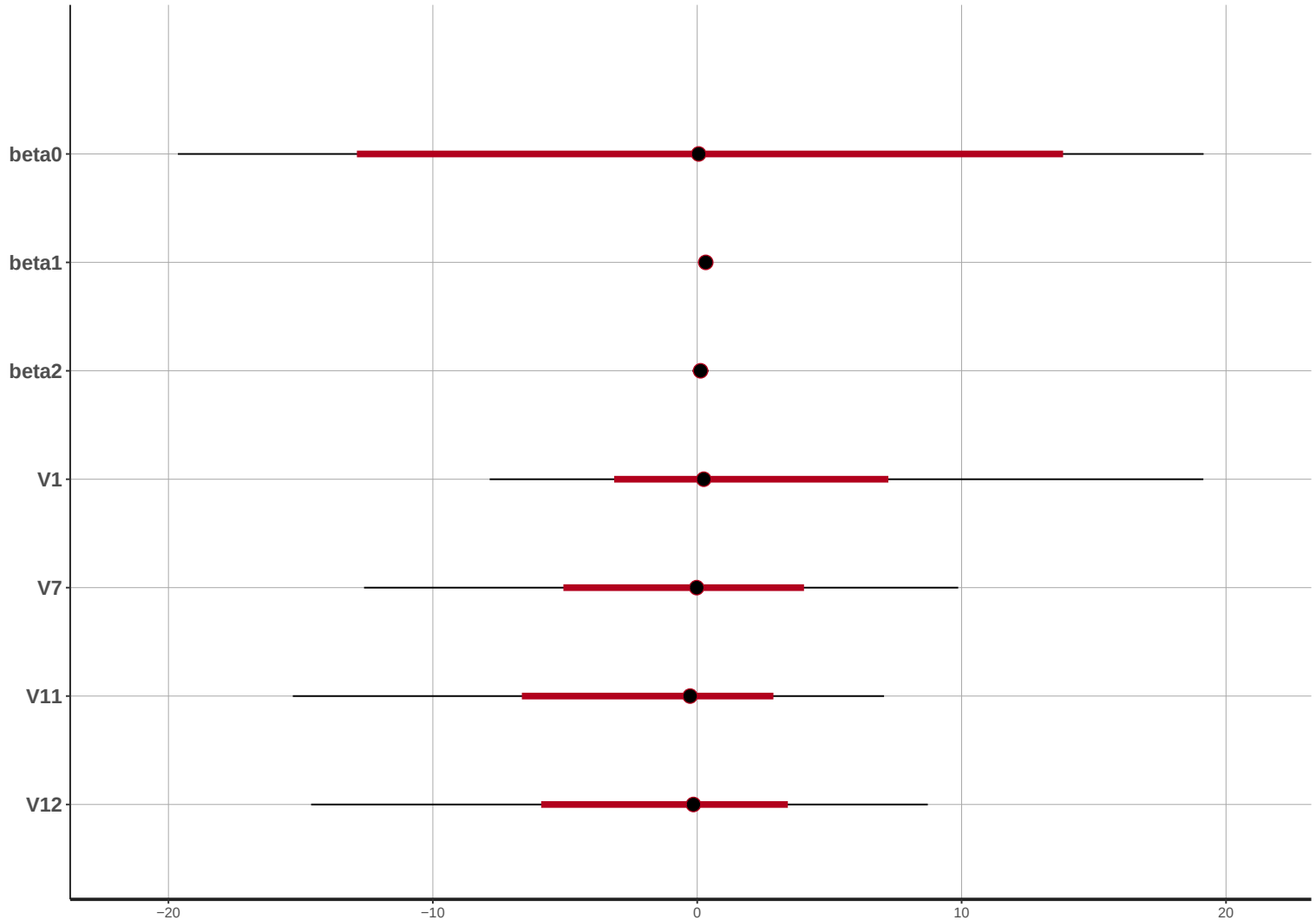


Figure: CI plot for parameter estimates from Model 6.

Parameter	mean	sd	se_mean	2.5%	50%	97.5%
theta[1]	120.00	8.97	0.11	105.11	119.22	140.81
theta[2]	120.19	8.12	0.05	103.33	120.37	135.53
theta[3]	117.24	8.31	0.06	98.66	117.63	131.69
theta[4]	122.49	8.65	0.08	107.48	121.69	142.15
theta[5]	128.03	9.90	0.22	111.06	127.31	150.45
theta[6]	112.69	12.19	0.13	88.20	112.85	136.54
theta[7]	116.49	7.51	0.05	101.37	116.49	131.50
theta[8]	125.82	8.01	0.07	111.55	125.20	143.73
theta[9]	115.31	11.46	0.09	92.63	115.43	137.96
theta[10]	123.95	7.88	0.08	107.09	124.27	138.53
theta[11]	120.20	7.07	0.05	105.22	120.36	133.43
theta[12]	125.93	9.77	0.19	106.77	126.05	145.07

Table: Estimates of θ_i's – the estimates of small area means – from Model 6

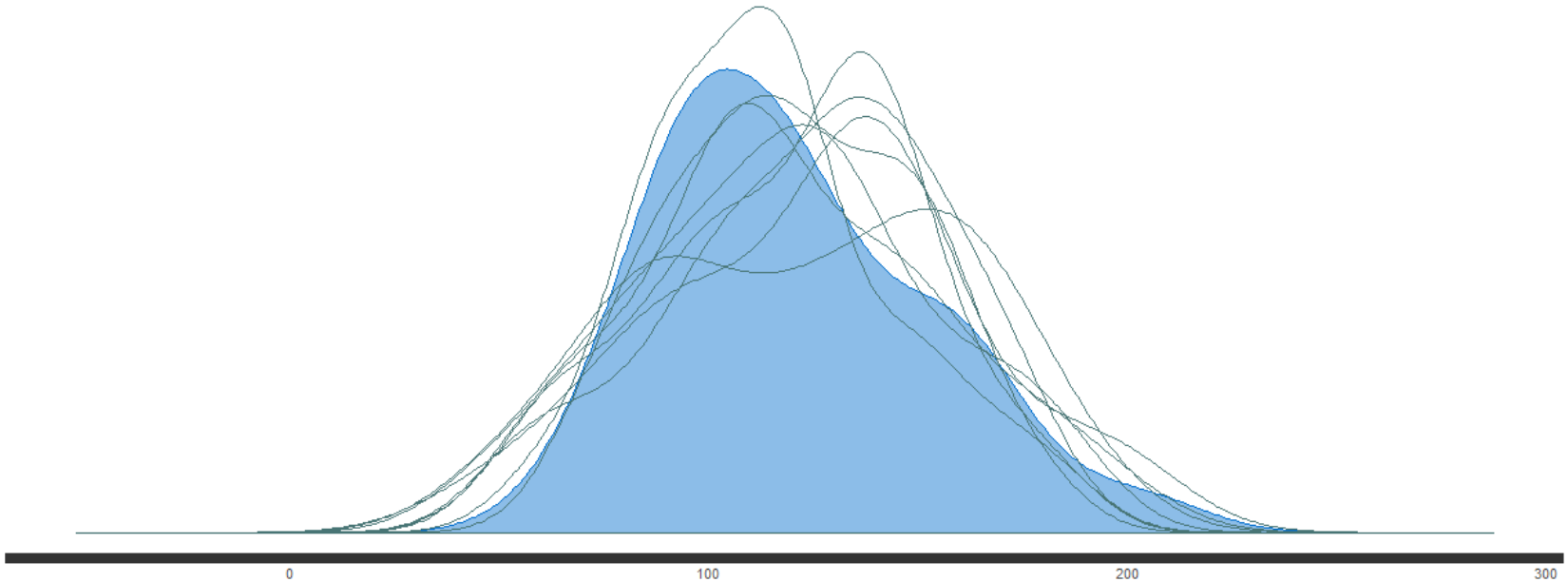


Figure: Posterior predictive simulations from Model 6. The shaded density is of the actual data.

Estimation of corn hectares: Hierarchical Bayes (HB)

- You and Rao (2003) carried out a hierarchical Bayesian analysis using BUGS program.
- Default flat prior for β and inverted Gamma priors for the variance components σ_v^2 and σ_e^2
- Based on the posterior predictive p-values and conditional predictive ordinate (CPO), nested error model with two auxiliary variables x_1 and x_2 was chosen, although the same model with x_1 only was also good.
- Single chain of length $D = 5000$ after discarding $d = 5000$ "burn-in" iterations.
- Psedo-Bayesian analysis was carried out by You and Rao (2003). This is a way to available weights and this method automatically calibrates the psuedo-HB estimates to direct survey regression estimate for higher level. the psuedo-HB

Different estimates of county corn areas

County	Samp. Size	Pop. Size	EB	P-EB	HB_6	HB_6^*	HB_4	P-HB
Cerro Gordo	1	545	122.2	120.5	122.2	120.0	121.60	120.6
Hamilton	1	566	126.3	125.2	126.1	120.19	121.42	125.2
Worth	1	394	106.2	106.4	108.1	117.24	116.55	107.5
Humboldt	2	424	108.0	107.4	109.5	122.19	120.10	108.4
Franklin	3	564	145.0	143.7	142.8	128.03	131.62	142.5
Pocahontas	3	570	112.6	111.5	111.2	112.69	104.59	110.6
Winnebago	3	402	112.4	112.1	113.8	116.49	118.28	113.2
Wright	3	567	122.1	121.3	122.0	125.82	124.15	121.1
Webster	4	687	115.8	115.0	114.5	115.31	107.77	114.2
Hancock	5	569	124.3	124.5	124.8	123.95	125.85	124.8
Kossuth	5	965	106.3	106.6	108.4	120.20	120.31	108.0
Hardin	5	556	143.6	143.5	142.2	125.93	131.24	142.3

Source: The table is compiled using Table 7.1 and Table 7.4 of Rao and Molina (2015)

Estimated standard errors of different estimates of county

County	Samp. size	se (EB)	se (P-EB)	se (Surv. Reg.)	se (HB_6)	se (HB_6^*)	se (HB_4)	se (P-HB)
Cerro Gordo	1	9.6	9.9	13.7	8.9	8.97	8.37	9.3
Hamilton	1	9.5	9.7	12.9	8.7	8.12	8.01	9.4
Worth	1	9.3	9.6	12.4	9.8	8.31	8.06	10.2
Humboldt	2	8.1	8.3	9.7	8.1	8.65	8.18	8.2
Franklin	3	6.5	6.6	7.1	7.3	9.90	8.50	7.4
Pocahontas	3	6.6	6.6	7.2	6.5	12.19	6.80	7.0
Winnebago	3	6.6	6.6	7.2	6.5	7.51	7.12	7.0
Wright	3	6.7	6.8	7.3	6.2	8.01	7.82	6.4
Webster	4	5.8	5.8	6.1	6.1	11.46	6.90	6.4
Hancock	5	5.3	5.4	5.7	5.2	7.88	7.58	5.4
Kossuth	5	5.2	5.3	5.5	6.3	7.07	7.06	6.6
Hardin	5	5.7	5.8	6.1	6.0	9.77	7.48	6.1

Source: The table is compiled using Table 7.1 and Table 7.4 of Rao and Molina (2015)

Let $y_i = (y_{i,s},\ y_{i,ns})$ with $y_{i,s}$ and $y_{i,ns}$ denote the sampled and non-sampled parts, respectively. We assume a hierarchical model for $y_i,\ i = 1, \cdots, m$ (e.g., nested error model on y_{ij} or in a logarithmic scale). Then the Bayes/BP of θ_i for the general case can be approximated as follows:

Step 1: Obtain L "census" files as $y^*_{i;l} = (y_{i,s}, y^*_{i,ns}),\ (l = 1, \ldots, L)$, where $y^*_{i,ns}$ is generated from the conditional distribution of $y_{i,ns}$ given $y_{i,s}$ with known hyperparameters.

Step 2: Bayes/BP of θ_i is approximated by $L^{-1} \sum_{l=1}^{L} g(y^*_{i;l})$.
To obtain, EB or HB change step 1. For EB, $y^*_{i,ns}$ is generated from the conditional distribution of $y_{i,ns}$ given $y_{i,s}$ with estimated hyperparameters. For HB. $y^*_{i,ns}$ is generated from the conditional distribution of $y_{i,ns}$ given $y_{i,s}$ under some prior assumptions on the hyperparameters.

Poverty Mapping

With the choice $g(y_{ij}) = \left(\frac{z-y_{ij}}{z}\right)^{\alpha} I(y_{ij} < z)$, we obtain a class of SGT poverty measures (Foster, Greer and Thornbecke, 1984):

$$F_{\alpha i}(y_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} \left(\frac{z - y_{ij}}{z}\right)^{\alpha} I(y_{ij} < z),$$

where

$$I(y_{ij} < z) = \begin{cases} 1 & \text{if } y_{ij} < z \ , \\ 0 & \text{otherwise,} \end{cases}$$

where α is a measure of the sensitivity of the index to poverty. In the above, y is a welfare variable (income, expenditure, etc.) of interest. For U.S. it is household income in the Small Area Income and Poverty Estimates (SAIPE) program. In the above, z denotes threshold under(s) which a unit is under poverty. In the U.S. SAIPE program, different thresholds are used depending on the household composition.

Poverty Incidence, Poverty Gap, Poverty Severity

Poverty Incidence ($\alpha = 0$):

$$F_{\alpha i}(y_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} I(y_{ij} < z)$$

Poverty Gap ($\alpha = 1$):

$$F_{\alpha i}(y_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} \left(\frac{z - y_{ij}}{z} \right) I(y_{ij} < z)$$

Poverty Severity ($\alpha = 2$):

$$F_{\alpha i}(y_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} \left(\frac{z - y_{ij}}{z} \right)^2 I(y_{ij} < z)$$

A few case studies

Case Study 1: Hierarchical Bayes (HB) estimation of smoking prevalence

Ha et al. (2014)

- We dichotomize the current smoking status into current smokers and current non-smokers. Current smokers are defined as those who had smoked at least 100 cigarettes in their lifetime and at the time of interview reported smoking every day or on some days.
- N_i: adult population size in state i ($i = 1, \ldots, m = 51$).
- y_{ij}: smoking status for an adult j in state i, ($i = 1, \ldots, m,\ j = 1, \ldots, N_i$).
- Parameters of interest: $\theta_i = \sum_{j=1}^{N_i} y_{ij}/N_i$ ($i = 1, \ldots, m$), the true proportion of adults in ith area.
- A direct survey-weighted estimate of θ_i:

$$\hat{\theta}_i = \frac{\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{j=1}^{n_i} w_{ij}},$$

where n_i is the number of responding adults) for area i and w_{ij} is the survey weight associated with unit j in area i ($i = 1, \ldots, m$, $j = 1, \ldots, n_i$).

Data Sources

- The National Health Interview Survey (NHIS): provides current smoking status and various survey design, geographic, demographic and health related variables.
- The American Community Survey (ACS): provides data on various geographic, demographic, and socio-economic variables from every U.S. county covering about 3 million addresses nationwide.

Two Alternatives to the Fay-Herriot model

The Normal-Logistic Model: (NL)

For $i = 1, \ldots, m$,

$$\text{Level 1 } (\textit{Sampling model}): \quad \hat{\theta}_i|\theta_i \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, \psi_i),$$

$$\text{Level 2 } (\textit{Linking model}): \quad \text{logit}(\theta_i)|\beta, A \overset{\text{ind}}{\sim} \mathcal{N}(x_i'\beta, A).$$

The Normal-Logistic Random Sampling Variance Model: (NL_{rs})

For $i = 1, \ldots, m$,

$$\text{Level 1: Sampling model}: \hat{\theta}_i|\theta_i \overset{\text{ind}}{\sim} \mathcal{N}\left(\theta_i, \psi_i = \frac{\theta_i(1-\theta_i)}{n_i}\text{DEFF}_i\right),$$

$$\text{Level 2: Linking model}: \text{logit}(\theta_i)|\beta, A \overset{\text{ind}}{\sim} \mathcal{N}(x_i'\beta, A),$$

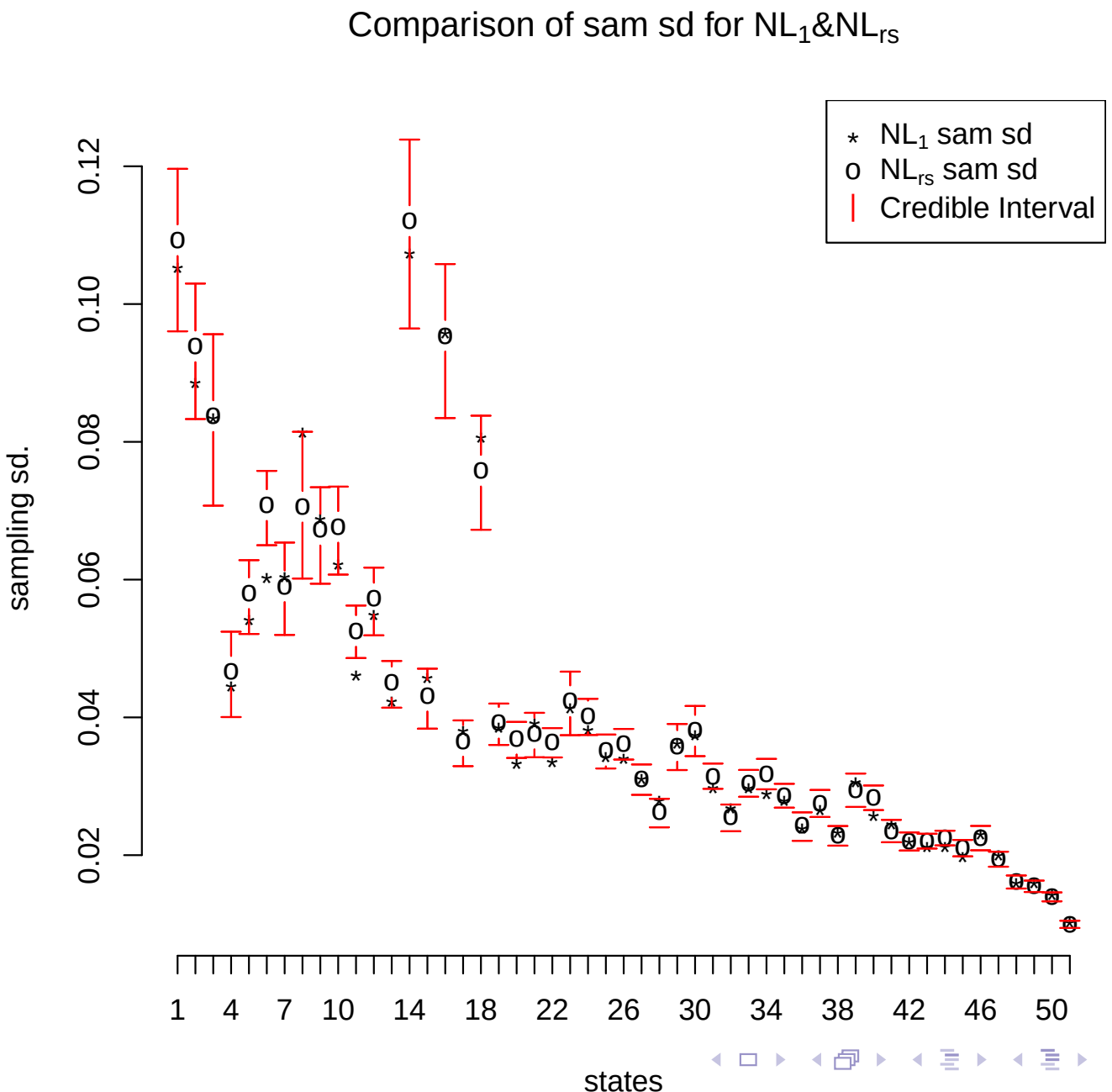
where DEFF_i is the true design effect.

Choices for ψ_i, DEFF_i, and x_i

- $\text{DEFF}_i = \widehat{\text{deff}}^{rgn}_{j(i)}$, a design-based estimator of the design effect for the jth census region, DEFF^{rgn}_j, in which ith state is located.
- $\psi_i = \frac{\hat{\theta}^{rgn}_{j(i)}(1-\hat{\theta}^{rgn}_{j(i)})}{n_i} \cdot \widehat{\text{deff}}^{rgn}_{j(i)}$
- Area level covariates:
  - Proportion of minority population
  - Proportion of education level
  - Proportion unemployment
  - Proportion senior population.

HB implementation through the MCMC

- Non-informative prior distributions for β and A,
- Purely Bayesian approach via Markov Chain Monte Carlo(MCMC) techniques, such as Metropolis Hastings(MH) algorithm, and Gibbs sampling.
- Three independent MCMC chains, Gelman et al.(2004)
- Convergence diagnostic tools:
  - Auto Correlation Function(ACF)
  - Geweke diagnostic criterion
  - Gelman and Rubin statistic, $\hat{R}$


Comparison of sam sd for NL1&NLrs
NL1 sam sd
NLrs sam sd
Credible Interval
sampling sd.
0.02
0.04
0.06
0.08
0.10
0.12
1
4
7
10
14
18
22
26
30
34
38
42
46
50
states

Model comparison

Relative error at the Census Regional level

- $$\left| \frac{\sum_{i \in j} w_i \hat{\theta}_i^{ps} - \hat{\theta}_{j,dsgn}}{\hat{\theta}_{j,dsgn}} \right|, j = 1, \ldots, 4$$
- $\hat{\theta}_{j,dsgn}$= the Census regional direct estimate
- w_i=survey weight for i
- $\hat{\theta}_i^{ps}$= posterior estimate at state i.

Cen. Rgn.	NL_1	NL_{rs}
Rgn 1: NE	0.0329	0.0268
Rgn 2: MW	0.0372	0.0335
Rgn 3: S	0.0542	0.0524
Rgn 4: W	0.0019	0.0010

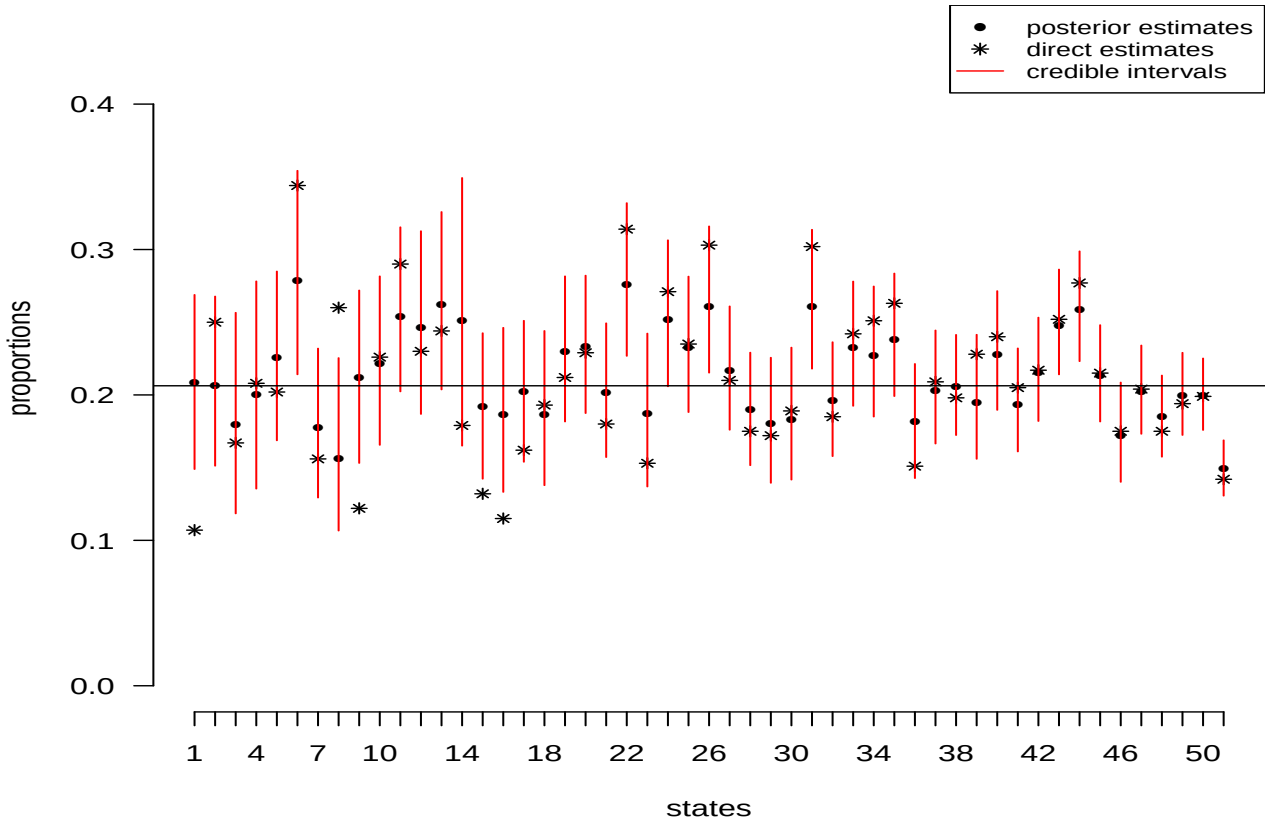
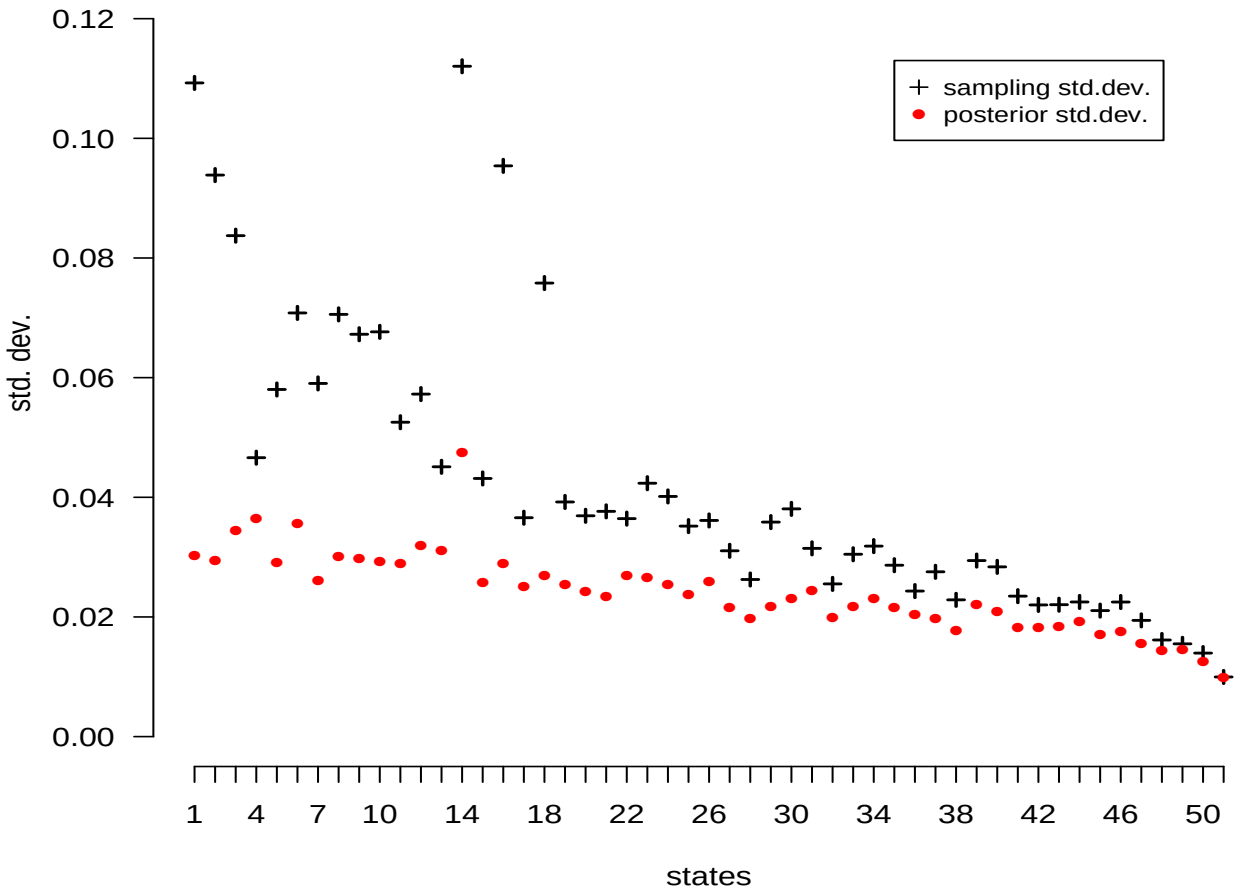


Figure: Coverage for model NL_{rs}

The vertical lines show the posterior credible intervals for each state. The credible intervals are larger for smaller states than bigger states. Most of the direct estimates are contained within the credible intervals.


0.12
0.10
0.08
0.06
0.04
0.02
0.00
std. dev.
sampling std.dev.
posterior std.dev.
1
4
7
10
14
18
22
26
30
34
38
42
46
50
states

Case Study 2: Use of Twitter Data in Estimating Race Distribution for Small Areas

Descriptive Model

Level 1: Individudal Model. The individual parameter $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{iK})$ is of ultimate interest, $\sum_{k=1}^{K} p_{ik} = 1$.

$$(y_{i1}, \ldots, y_{iK}) | p_{i1}, \ldots, p_{iK} \sim \textit{Multinomial}(n_i, p_{i1}, \ldots, p_{iK}),$$

Level 2: Structural Model. Given the structural parameters,

$$(p_{i1}, \ldots, p_{iK}) | \boldsymbol{\beta}, r \sim \textit{Dirichlet}(rp_{i1}^E, \ldots, rp_{iK}^E),$$

where $p_{ik}^E = \frac{e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k}}{1 + \sum_{j=1}^{K-1} e^{\boldsymbol{x}_i' \boldsymbol{\beta}_j}}$ for $k = 1, \ldots, K-1$ and

$p_{iK}^E = \frac{1}{1 + \sum_{j=1}^{K-1} e^{\boldsymbol{x}_i' \boldsymbol{\beta}_j}}$ so that $\sum_{k=1}^{K} p_{ik}^E = 1$.

Level 3: Distributions on the Structural Parameters.

$$\boldsymbol{\beta}_k \sim \textit{Uniform on } \mathbb{R}^q$$

for $k = 1, \ldots, K-1$, and

$$1/r \sim \textit{Uniform}(0, \infty).$$

Data Example

- Data Collection
  Twitter Streaming API: This dataset contains 161,771,878 Twitter messages sent by 3,670,604 active Twitter users between July 10, 2017 and October 20, 2017 in the continental United States. Contains information about the user (geotag and self-reported name).

- Individual Race Distribution
  Extract the self-reported name from each tweet and use the last name to infer about the race distribution of the Twitter user using Census Bureau's surname list. For example, the surname 'Taylor' is 65.38% Caucasian (non-hispanic), 28.42% African-American, 0.56% Asian or Pacific Islander and 2.46% Hispanic and 3.18% of being other races.

Data Example - continued

- Location
  Location information can be inferred from the geotag contained in each tweet. Feed the geotag to Bing Maps API to get the longitude and lattitude of each user. Use the PUMA shapefile to assign a PUMA to each Twitter user.

- Race Counts
  Let s_i denote the set of observations in the i^{th} PUMA and d_{jk} denote the probability of the race k for the j^{th} Twitter user in s_i. Then the count for race k in the i^{th} PUMA is y_{ik} and is calculated as follows

$$y_{ik} = \sum_{j \in s_i} d_{jk}.$$

Comparison of ADM and MCMC: Data Example

	Data						MCMC $\hat{\beta} = \begin{pmatrix} 2.62 & -0.21 \\ 1.10 & -0.19 \\ 0.52 & -0.02 \\ 1.54 & 0.07 \end{pmatrix}, \hat{r} = 43.975$					ADM $\hat{\beta} = \begin{pmatrix} 2.54 & -0.20 \\ 1.05 & -0.18 \\ 0.50 & -0.02 \\ 1.47 & 0.07 \end{pmatrix}, \hat{r} = 42.660$				
obs i	y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}	x_{i1}	$\hat{p}_{i1}$	$\hat{p}_{i2}$	$\hat{p}_{i3}$	$\hat{p}_{i4}$	$\hat{p}_{i5}$	$\hat{p}_{i1}$	$\hat{p}_{i2}$	$\hat{p}_{i3}$	$\hat{p}_{i4}$	$\hat{p}_{i5}$
1	289	62	45	108	19	1	0.55 (0.02)	0.12 (0.01)	0.09 (0.01)	0.21 (0.02)	0.04 (0.01)	0.55 (0.02)	0.12 (0.01)	0.09 (0.01)	0.21 (0.02)	0.04 (0.01)
2	261	65	46	187	19	1	0.46 (0.02)	0.11 (0.01)	0.08 (0.01)	0.32 (0.02)	0.03 (0.01)	0.46 (0.02)	0.11 (0.01)	0.08 (0.01)	0.32 (0.01)	0.03 (0.01)
3	2	0	1	4	1	1	0.47 (0.09)	0.10 (0.05)	0.09 (0.05)	0.28 (0.08)	0.06 (0.04)	0.47 (0.08)	0.10 (0.05)	0.09 (0.04)	0.28 (0.08)	0.06 (0.04)
4	233	45	19	58	13	0	0.62 (0.02)	0.12 (0.02)	0.05 (0.01)	0.16 (0.02)	0.04 (0.01)	0.63 (0.02)	0.12 (0.02)	0.05 (0.01)	0.16 (0.02)	0.04 (0.01)
5	172	41	10	43	9	0	0.62 (0.03)	0.15 (0.02)	0.04 (0.01)	0.16 (0.02)	0.03 (0.01)	0.62 (0.03)	0.15 (0.02)	0.04 (0.01)	0.16 (0.02)	0.03 (0.01)

Monte Carlo Simulation Results

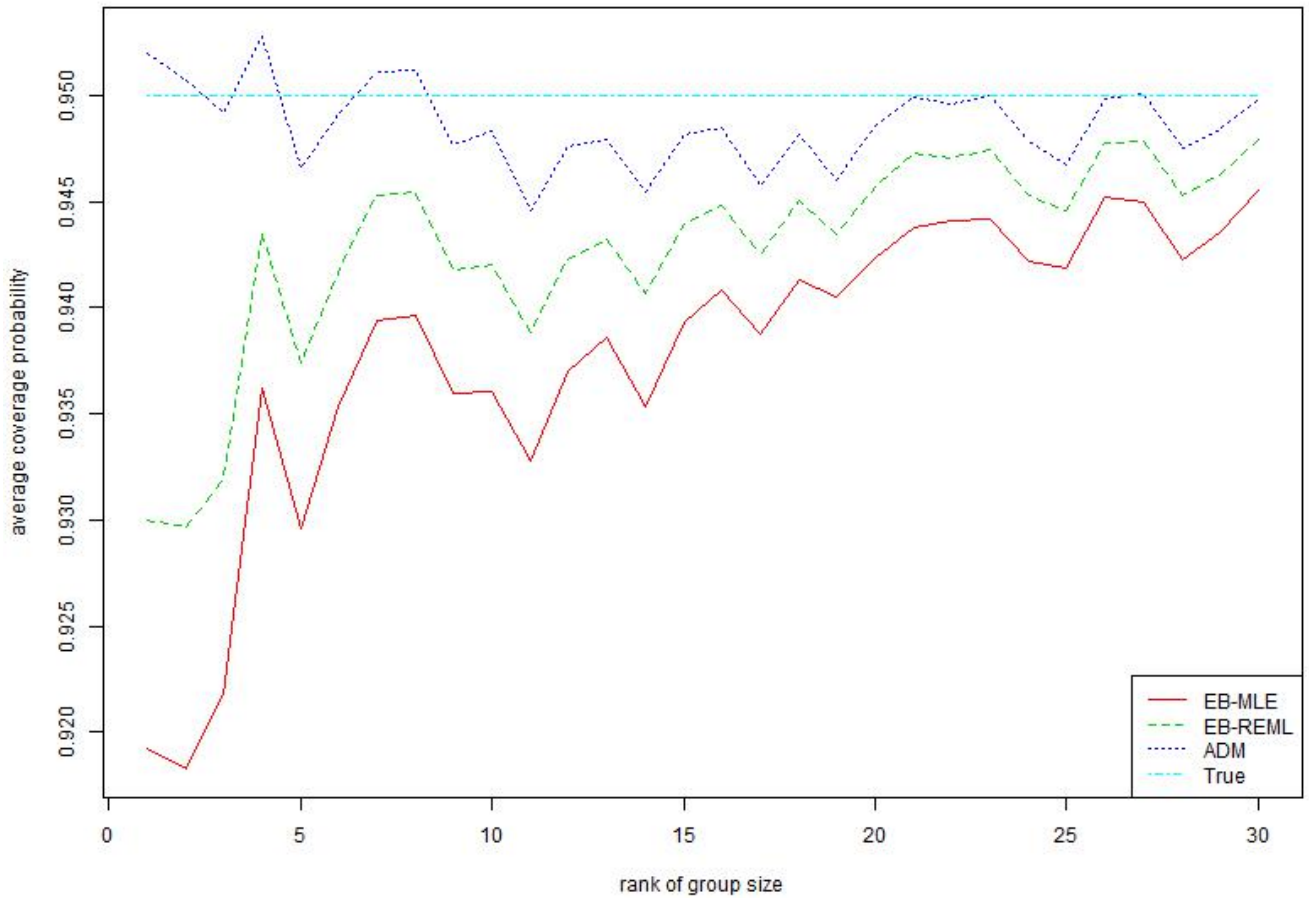


Figure: Average coverage rate vs group index, N=30, K=5, q=2, 100 samples of the dataset. Group size increases from 10 to 139 from group 1 to group 30. The total average coverage rates for EB-MLE, EB-REML and ADM are 0.938,0.943 and 0.949, respectively.

Summary

1. Computation speed: 180 times faster than MCMC without sacrificing the accuracy of the estimates.
2. Estimates are the same when applied to the same dataset using the same model.
3. Capability to handle non-integer counts without rounding.
4. Eliminates the ill-behavior in MLE and REML estimates and corrects the bias in r estimate
5. Better operating characteristics

References

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, 28-36.

Bell, W. R., and Franco, C. (2017). Small Area Estimation – State Poverty Rate Model Research Data Files. Available at `https://www.census.gov/srd/csrmreports/byyear.html` [accessed October 22, 2018]

Bell, W. R, Basel W. W., Cruse C. S., Dalzell L., Maples J. J., O'Hara, B. J, Powers D. S. (2007). Use of ACS Data to Produce SAIPE Model-Based Estimates of Poverty for Counties. Available at `https://www.census.gov/content/dam/Census/library/working-papers/2007/demo/bellreport.pdf`.

Bell, W. R., Basel W. W., Maples, J. J. (2016). An Overview of the U.S. Census Bureau's Small Area Income and Poverty Estimates Program. In M. Pratesi (Ed.) *Analysis of Poverty Data by Small Area Estimation* (pp. 349-377). West Sussex: Wiley & Sons, Inc.

Brackstone, G.J. (1987) Small area data: policy issues and technical challenges. In *Small Area Statistics* (R.L. Platek et al eds.), 3-20, Wiley, New York.

Cao, L. (2018). Adjustment for density method to estimate random effects in hierarchical Bayes model, PhD dissertation, University of Maryland, College Park.

Casas-Cordero, C., Encina, J. and Lahiri, P. (2015). Poverty mapping for the Chilean comunas. In *Analysis of Poverty Data by Small Area Estimation* (M. Pratesi, ed.). Wiley, New York.

Elbers, C., J. O. Lanjouw & P. Lanjouw (2003). Micro-level estimation of poverty and inequality. *Econometrica,* 71, 355-364.

Erciulescu, A. L. Franco, C. and Lahiri, P. (2018), Use of Administrative Records in Small Area Estimation, In Administrative records for survey methodology, Wiley Series in Survey Methodology, eds. Asaph Young Chun and Michael Larsen, forthcoming.

Fay, R. E. III and Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.* 74 269277. MR0548019

Ghosh, M. and Rao, J.N.K. (1994), Small area estimation: an appraisal, *Statistical Science*, 9, 55-76.

Ha, N. S., Lahiri, P. and Parsons, V. (2014). Methods and results for small area estimation using smoking data from the 2008 National Health Interview Survey, *Statistics in Medicine*. **33**. 22.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953), *Sample Survey Methods and Theory*, 2 Volumes. New York: John Wiley and Sons.

Jiang, J., and Lahiri, P. (2006), Mixed model prediction and small area estimation, Editor's invited discussion paper, *Test,* Vol. 15, 1, 1-96.

Lahiri, P. (2003), On the impact of bootstrap in survey sampling and small-area estimation, *Statistical Science*, Vol. 18, 199-210.

Lahiri, P. (2003a), A review of empirical best linear unbiased prediction for the Fay-Herriot small- area model, *The Philippine Statistician,* Vol 52, nos. 1-4, 1-15.

Lahiri, P. (2018), SAE Methods in Official Data Production for Policy Making (lecture video) `http://sampleu.ec.unipi.it/day-2-jean-monnet-chair-sampleu-workshop-9-may-2018/`

Lahiri, P. and Suntornchost, J. (2015) Variable Selection for a Regression model when dependent variable is subject to measurement errors, *Sankhya*, Series B. DOI 10.1007/s13571-015-0096-0

Lahiri, P. and Suntornchost, J. (2018) A general Bayesian approach for solving different inferential problems in poverty research for small areas, arxiv paper.

Liu, B., Lahiri, P. and Kalton, G. (2014). Hierarchical Bayes Modeling of Survey-Weighted Small Area Proportions. *Survey Methodology.* 40. 1-13.

Molina, I. and Marhuenda, Y. (2015), sae: An R Package for Small Area Estimation, The R Journal Vol. 7/1, June 2015.

National Research Council (2000b). Small Area Income and Poverty Estimates: Priorities for 2000 and Beyond (eds Citro C.F. and Kalton G.), Panel on Estimates of Poverty for Small Geographic Areas, Committee on National Statistics, Washington, DC: National Academy Press.

Otto M.C. and Bell W.R. (1995). Sampling error modelling of poverty and income statistics for states. *Proceedings of the American Statistical Association*, Social Statistics Section, 160165. Avaiable online at `https://www.census.gov/library/working-papers/1995/demo/otto-01.html` [accessed October 25 2018]

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science* 28, 4068.

Rao, J. N. K. and Molina, I. (2015). Small Area Estimation, 2nd ed. Wiley, Hoboken, NJ. MR3380626.

Tzavidis, N., Zhang, L.C., Pfeffermann, D. and Lahiri, P. (2017), Special issue on small area estimation, Journal of the Royal Statistical Society, Series A.

Upcoming Events

ISI SATELLITE CONFERENCE ON CURRENT TRENDS IN SURVEY STATISTICS, SINGAPORE, AUGUST 13-16, 2019
`https://ims.nus.edu.sg/orgsites/2019data/`

WORKSHOP ON STATISTICAL DATA INTEGRATION, SINGAPORE, AUGUST 5-8, 2019

Contact Information

Partha Lahiri
Joint Program in Survey Methodology
Department of Mathematics
1218 Lefrak Hall
University of Maryland
College Park, MD 20742
Email: plahiri@umd.edu

Thank You!