National Cancer Institute

Small Area Estimation for Measures Related to Tobacco Use and Policies Using the Tobacco Use Supplement to the Current Population Survey

Benmei Liu, National Cancer Institute

Isaac Dompreh, U.S. Census Bureau

Presented at the FCSM
March 8, 2018
Washington, DC

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health

Disclaimer

This presentation is intended 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 authors and not necessarily those of the National Cancer Institute and the U.S. Census Bureau.

Outline

- Background of TUS-CPS and research goal
- SAE models and inferences
- Simulation studies to decide on a model and inference approach
- Implementation of the chosen model and evaluation of the model-based estimates
- Demonstration of final results
- Summary and discussion

TUS-CPS Background

- NCI sponsored survey (partially sponsored by CDC, FDA)
- Key source of U.S. national and state level data on tobacco use and tobacco control policy
- Supplement to the BLS' Current Population Survey (CPS) conducted by the Census Bureau
  - uses a national complex probability address-based sample of households
  - is conducted monthly, uses panel design for sampling efficiency
  - detailed stats on demography, labor force & unemployment
- TUS fielded about every two to four years since 1992
  - 240,000 civilian individuals aged 15+(now 18+)
  - 70% by phone & 30% in-person visit interview
  - translated into Spanish

http://riskfactor.cancer.gov/studies/tus-cps

Research Goal

- Model-based county level estimates for the following key measures (for population aged 18+):
  1. Percent of population currently smoking
  2. Percent of population that has ever smoked
  3. Percent of population that has quit for 24+ hours, among those who have smoked within the past year
  4. Percent of population governed by a smoke-free workplace policy
     - (Workplace has an official smoking policy: Smoking Not allowed in ANY public areas and work areas)
  5. Percent of population governed by a smoke–free home rule
     - (No one is allowed to smoke anywhere INSIDE YOUR HOME)

✓ Important to the Tobacco Control Research Branch

Notations

y_{ik} : a binary response for unit *k* in county i;

s_i : the set of sampled units in county i;

n_i : the sample size in county i;

w_{ik} : the sampling weight for unit *k* in county i;

$\mathbf{x}_i$: the vector of auxiliary variables;

$$k = 1, \ ..., \ N_i; \quad i = 1, \ ..., \ m$$

- Parameters of interest are the population proportions:

$$P_i = \sum_{k=1}^{N_i} y_{ik} / N_i$$

Direct Estimates of P_i and Associated Variances

- Direct estimates (design-unbiased):

$$p_{iw} = \frac{\sum_{k \in s_i} w_{ik} y_{ik}}{\sum_{k \in s_i} w_{ik}}, i = 1, \ldots, m$$

- Variances of the direct estimates:

$$VAR_{st}(p_{iw}) = \frac{P_i(1-P_i)}{n_i} * DEFF_i, \; i = 1, \ldots, m.$$

Where $DEFF_i$ is the design effect reflecting the complex design (Kish 1965).

- Problem of p_{iw}: Variance too large (imprecise estimates) for small sample sizes n_i
- Small area estimation techniques to address imprecise estimates

Commonly Used Area Level Model: Fay-Herriot Model

The well known Fay-Herriot model (Fay & Herriot 1979):

- Sampling model: $p_{iw}|P_i \sim N(P_i,\ D_i)$;
  - D_i is the sampling variance and is assumed known
- Linking model: $P_i = X_i'\beta + v_i$; where $v_i \sim N(0, A)$;

- Several transformations proposed to stabilize sampling variance D_i:
  - Fay and Herriot (1979): $\hat{\theta}_i = \log(p_{iw})$
  - Efron and Morris (1975): $\hat{\theta}_i = \sqrt{n_i}\arcsin(2p_{iw} - 1)$
  - Carter and Rolph (1974): $\hat{\theta}_i = \arcsin\left(\sqrt{p_{iw}}\right)$

Fay-Herriot Model with C&R Arcsin Transformation

Let $z_i = \arcsin\left(\sqrt{p_{iw}}\right)$; (Carter & Rolph, 1974 JASA)

- Sampling model: $z_i|\theta_i \sim N\left(\theta_i, \frac{DEFF_i}{4n_i}\right)$;
- Linking model: $\theta_i = X_i'\beta + v_i$; where $v_i \sim N(0, A)$

→ Goal: To estimate $P_i = sin^2(\theta_i)$

!! Extensive simulation study conducted to decide the model to use

SAE Model: Empirical Bayes Predictor (EBP)

Empirical Bayes estimator for P_i, $i = 1, \ldots, m$:

$$\hat{P}_i^{EB} = (1 - \hat{b}_i)p_{iw} + \hat{B}_i X_i' \hat{\beta}, \text{ where } \hat{B}_i = \frac{D_i}{\hat{A} + D_i}.$$

- $\hat{\beta}$: can be obtained using maximum likelihood estimation (MLE)
- $\hat{A}$: restricted maximum likelihood estimation (REML)
- MSE estimation of $\hat{P}_i^{EB}$ using Delta method (Datta and Lahiri 2000):

  MSE($\hat{P}_i^{EB}$)=$g_{1i}(\hat{A}, \hat{\beta}) + g_{2i}(\hat{A}, \hat{\beta}) + 2g_{3i}(\hat{A}, \hat{\beta})$, normal confidence interval can be computed using $\hat{P}_i^{EB}$ and MSE($\hat{P}_i^{EB}$).
- Parametric bootstrap prediction interval (Chatterjee, Lahiri and Li 2008)
  - Step 1: Generate R bootstrap samples using distributions: $P_i^* \sim N(x_i'\hat{\beta}, \hat{A})$ and $p_{iw}^* \mid P_i^* \sim N(P_i^*, D_i), i = 1, \ldots, m, \hat{\beta}$ and $\hat{A}$ obtained from the original sample.
  - Step 2: For each of the R bootstrap samples, obtain $\hat{A}^*$,$\hat{B}_i^*$, and $\hat{\beta}^*$
  - Step 3: For each bootstrap sample, compute $t_i^* = (P_i^* - \hat{P}_i^{EB*}) / \sqrt{D_i(1 - \hat{B}_i^*)}$
  - Step 4: Locate the two equal-tail $\alpha/2$ cut-off points (t_1, t_2) using the B pivot values computed in step 3.
  - Step 5: Construct the parametric bootstrap prediction interval for small area i as follow:

$$PI_i = \left[\hat{P}_i^{EB} + t_1\sqrt{D_i(1 - \hat{B}_i)}, \hat{P}_i^{EB} + t_2\sqrt{D_i(1 - \hat{B}_i)}\right]$$

Hierarchical Bayes (HB) Estimator

- Prior assumptions:
  - Flat prior for β, i.e., $f(\beta) \propto 1$
  - Uniform prior for A, i.e., $A \sim unif(0, L)$, where L is a large number.
- Multiple (3) independents chains for each run, Burn-ins of 10,000 samples and 10,000 after burn-in, thinning=2
- Posterior mean and percentiles obtained from the 15,000 MCMC samples (3 chains of 5,000 independent samples)
- 95% credible interval for $\hat{P}_i^{HB}$ is $\left(\hat{P}_{i,2.5}^{HB}, \hat{P}_{i,97.5}^{HB}\right)$

Model-based Simulation study: Data Generation

- Step 1: Obtain β and A using the 2002 Natality public-use data;
  - Parameter of interest is: $P_i = P(weight\ at\ birth < 3{,}345\ grams) \in [40.2\%, 58.5\%]$, where 3,345 grams was the 2002 national median birthweight.
  - Five auxiliary variables were chosen (percent of births with mother being non-Hispanic, percent of births being first live child in family, etc)
  - Fit the model $\hat{P}_i = x_i'\beta + v_i$; where $v_i \sim N(0, A)$ to obtain estimate of β and A.
- Step 2: Generate one set of θ_i from $\theta_i \sim N(x_i'\hat{\beta}, \hat{A})$, $i = 1, \ldots, 51$.
- Step 3: Generate 1,000 sets of observed data p_{iw} using three approaches:
  - Approach a: Generate p_{iw}based on level 1 of the Fay-Herriot model $p_{iw} \sim N(\theta_i, D_i)$
  - Approach b: Generate binary y_{ij} using Poisson distribution, $y_{ij} \sim Pois\left(\frac{n_i\theta_i}{DEFF_i}\right)$ and then compute $p_{iw} = \frac{\sum_j y_{ij}}{n_i/DEFF_i}$
  - Approach c: Generate binary y_{ij} using binomial distribution $y_{ij} \sim Bin\left(int\left(\frac{n_i}{DEFF_i}\right), \theta_i\right)$ and then compute $p_{iw} = \frac{\sum_j y_{ij}}{int(n_i/DEFF_i)}$

Model-based Simulation Study: SAE Models and Estimators

- Two modeling approaches (with five covariates)
  - Fay-Herriot model
  - Fay-Herriot model with arcsin transformation
- Three estimators:
  - EBP Parametric Bootstrap
  - EBP-REML-Delta method
  - HB method

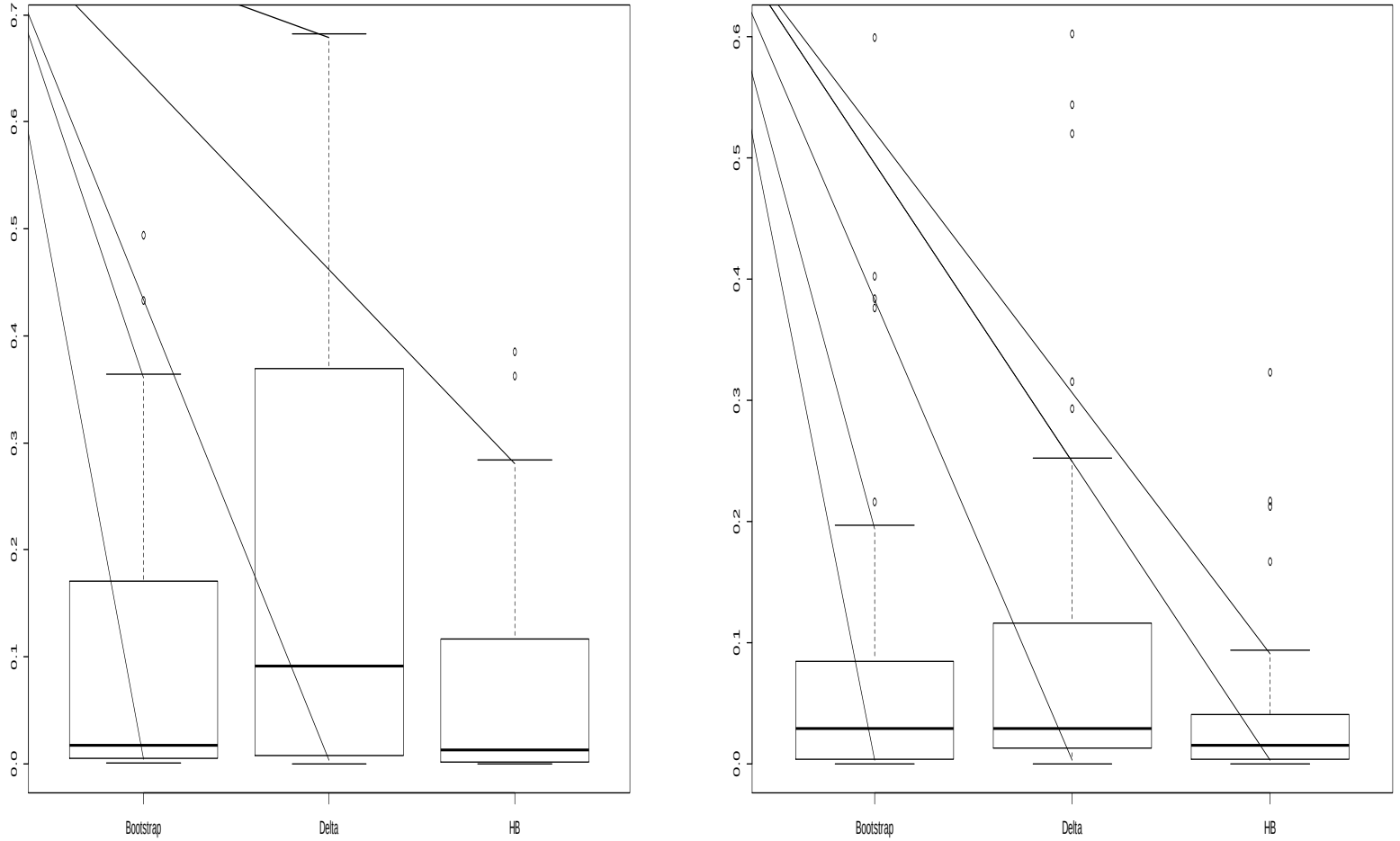
Simulation Results: Summary on 95% Confidence Interval Non-Coverage Rates and MC Errors

Data generation approach	State sample size n	Fay-Herriot Model			Fay-Herriot with arcsin transformation		
		EBP Parametric Bootstrap	EBP REML-Delta	HB	EBP Parametric Bootstrap	EBP REML-Delta	HB
Approach A	overall	4.9 (0.09)	6.9 (0.1)	3.9 (0.08)	4.5 (0.09)	6.4 (0.1)	4.1 (0.09)
	50<=n<60 (24 states)	4.9 (0.13)	6.2 (0.15)	3.0 (0.11)	4.6 (0.13)	5.8 (0.14)	3.6 (0.12)
	60<=n<100 (15 states)	5.6 (0.18)	8.5 (0.21)	4.7 (0.16)	5.2 (0.17)	7.9 (0.2)	4.5 (0.16)
	100<=n<=690 (12 states)	3.9 (0.17)	6.2 (0.2)	4.7 (0.19)	3.6 (0.16)	5.8 (0.2)	4.7 (0.19)
Approach B	overall	4.5 (0.09)	6.3 (0.1)	3.7 (0.08)	4.2 (0.08)	6 (0.1)	3.6 (0.08)
	50<=n<60 (24 states)	3.2 (0.11)	3.9 (0.12)	2 (0.09)	3 (0.11)	3.7 (0.12)	2.1 (0.09)
	60<=n<100 (15 states)	4.9 (0.17)	7.3 (0.2)	3.6 (0.15)	4.5 (0.16)	6.7 (0.19)	3.2 (0.14)
	100<=n<=690 (12 states)	6.5 (0.2)	10.1 (0.24)	7.1 (0.22)	6.1 (0.2)	9.6 (0.24)	7 (0.22)
Approach C	overall	6.7 (0.11)	8.6 (0.12)	9.3 (0.13)	4.2 (0.08)	10 (0.13)	10.6 (0.14)
	50<=n<60 (24 states)	6.0 (0.15)	7.2 (0.17)	7.7 (0.17)	3.0 (0.11)	8.6 (0.18)	9.2 (0.19)
	60<=n<100 (15 states)	6.9 (0.2)	8.5 (0.23)	9.2 (0.23)	4.5 (0.16)	10.2 (0.24)	10.7 (0.25)
	100<=n<=690 (12 states)	7.7 (0.24)	11.5 (0.29)	12.5 (0.3)	6.1 (0.2)	12.5 (0.3)	13.3 (0.31)

Approach A: Generate p_{iw} based on the Level 1 of the Fay-Herriot model; **Approach B**: Generate y_i using poisson distribution; **Approach C**: Generate y_i using Binomial distribution

Design-based Simulation and Results

- Boxplots of confidence interval non-coverage rates (without-covariate vs five-covariates)



- Sampling frame: 2002 Natality public-use data file;
- Parameter of interest: state level percent of births with weight less than the national median birth weights;
- Repeatedly draw 1000 sets of samples using stratified SRS within states;
- Fit Fay-Herriot model and calculate three estimators

Conclusions from the Simulation Studies

- Design-based study
  - Covariates improved the coverage rates
  - None of the three approaches produced coverage rates close to the nominal rates
  - HB estimator produced more conservative intervals.
- Model-based study
  - Parametric bootstrap approach gave the best coverage rates
  - Parametric bootstrap approach produced coverage rates close to the nominal rates only when data generated from Fay-Herriot model
  - HB approach gave the next best coverage property with often conservative credible/confidence intervals.
  - Not much difference between Fay-Herriot and Fay-Herriot with arcsin transformation.

➢ Fay-Herriot with arcsin transformation and HB approach were chosen for practical reason

TUS-CPS: Estimate Design Effects

Kish's DEFF formula (Kish 1987)

$$DEFF_{kish} = n \frac{\Sigma w_i^2 n_i}{\left(\Sigma w_i n_i\right)^2} [1 + (\bar{b} - 1)\rho]$$

where n_i and w_i denote the # of observations and the weight attached to the ith weight class, $i = 1, \ldots, m$. $n = \sum_i n_{i.}$ $\bar{b}$ is the average cluster size; ρ is the intra-class correlation coefficient.

Steps:

- Use the national DEFF to estimate ρ
- Plug ρ into the Kish's formula to estimate the state level DEFF
- Use the state level DEFF to estimate the county level DEFF

Auxiliary Variables

- The pool of auxiliary variables include:
  - 30 county-level demographic & socio-economic variables obtained from the five-year average of ACS (2005-2009, 2008-2012, , Census 2000 & 2010, and other administrative records;
  - 5 state level tobacco policy data (cigarette taxes, clean air laws, tobacco control funding, Medicaid Coverage for Tobacco-Related Treatment, year in which Quitline service was established)
- Classical model selection procedures are applied to reduce the number of auxiliary variables for each outcome
- Tested forcing in several strong unit level covariates: only worked for current smoking and smoking cessation.

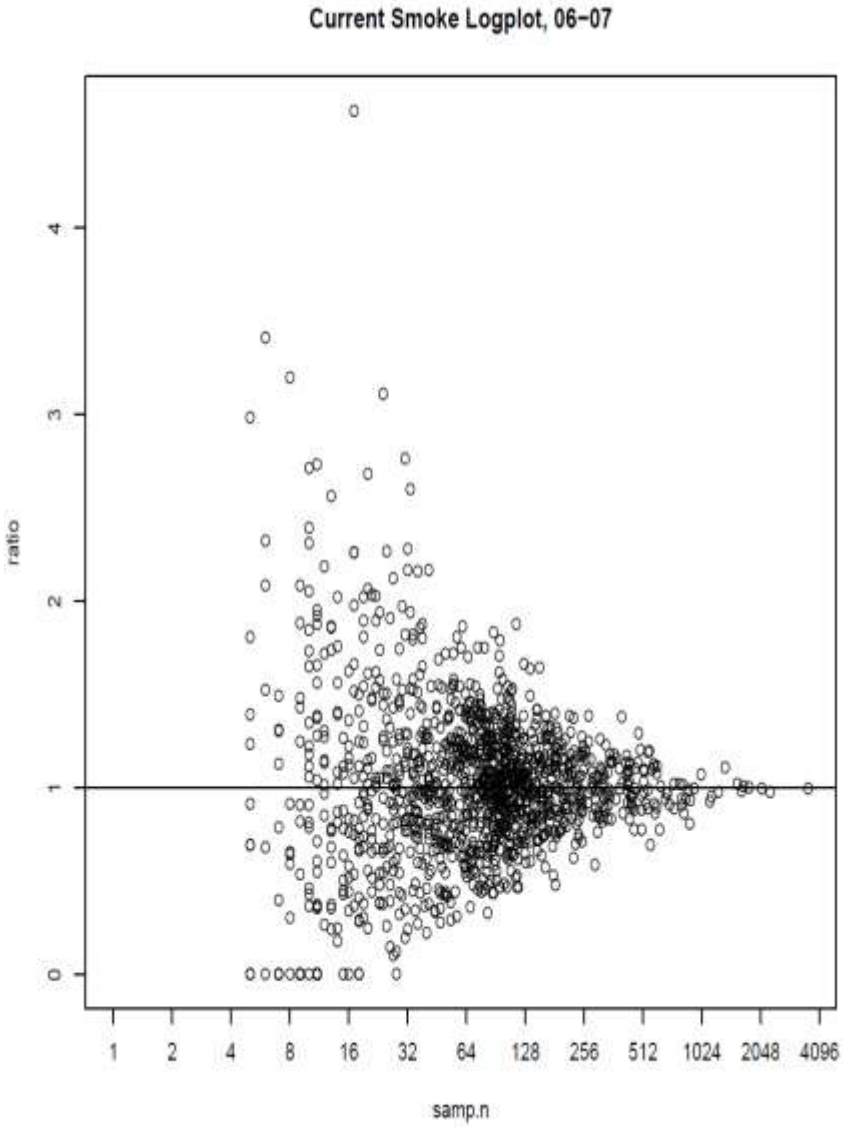
Statistical Inference and Model Diagnosis

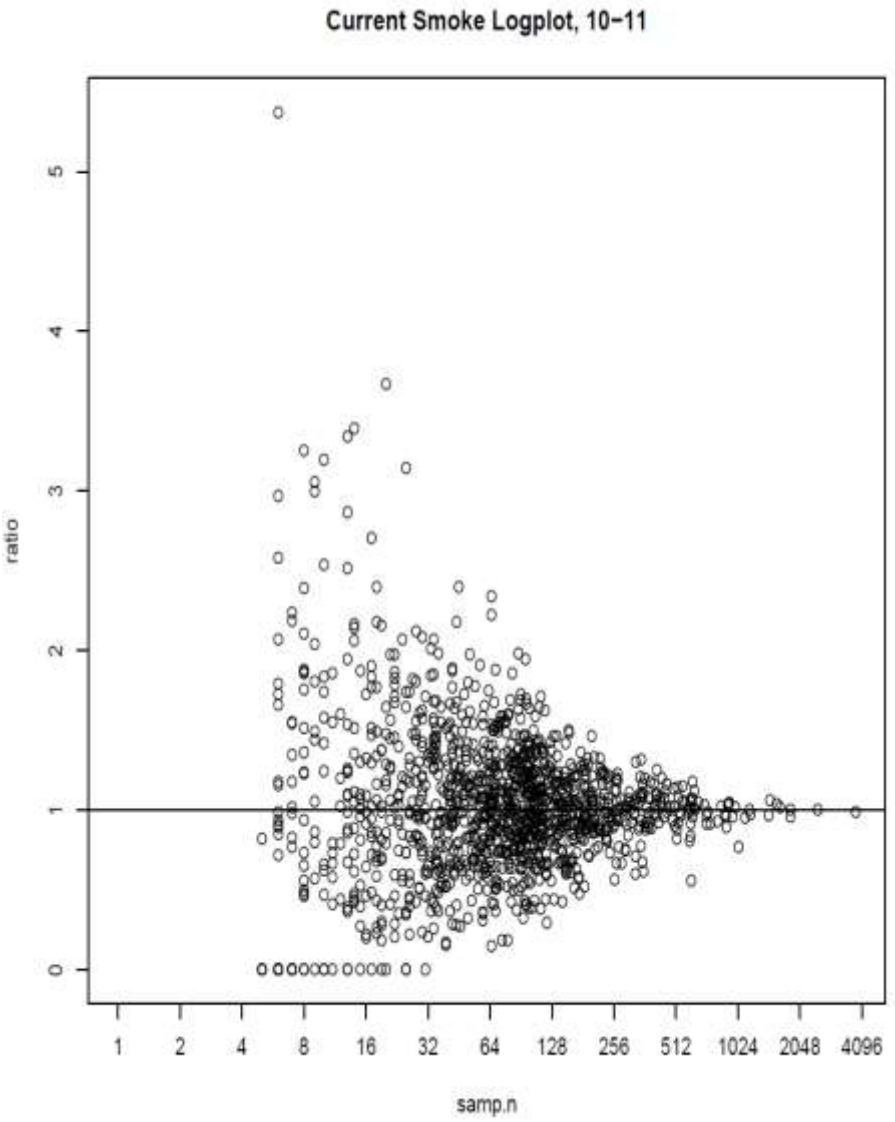
- HB approach through Markov Chain Monte Carlo (MCMC) methods are used to estimate the parameters of the statistical models.

- Extensive model selection and model diagnosis procedures are used to select the final models and assess the goodness of fit for each model.
  - Deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002)
  - Gelman and Rubin's Potential Scale Reduction Factor $\hat{R}$
  - Check the overall fit of the proposed model using method of posterior predictive p –value (Gelman & Meng 1996)

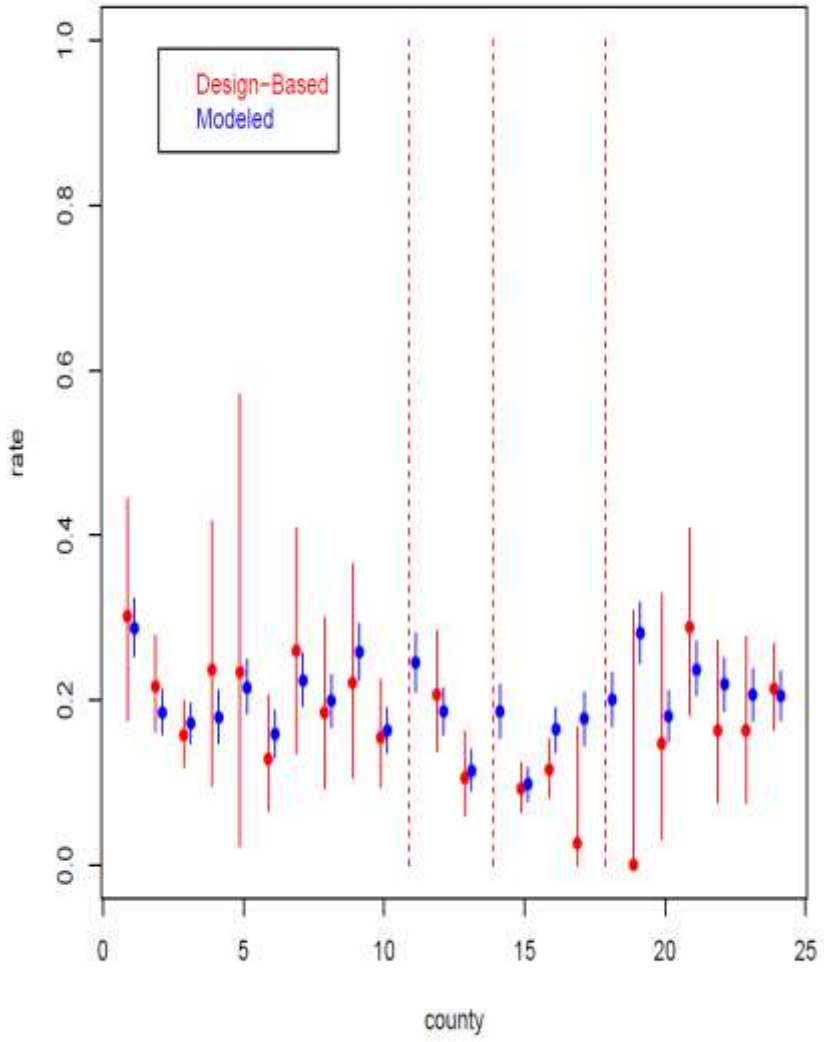
- Modeled estimates are also compared to the available direct estimates. The ratio of the two is expected to converge to 1 as the sample size gets larger.

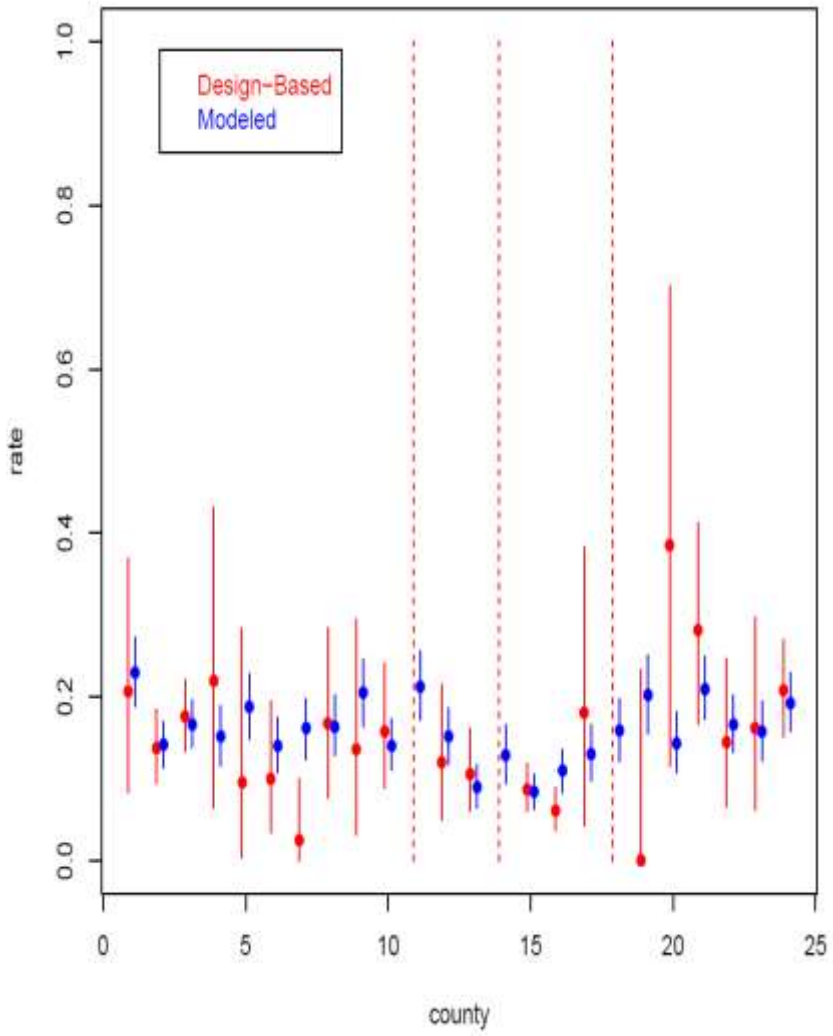
Ratio of the Direct Over the Modeled Estimates for the Current Smoking Prevalence





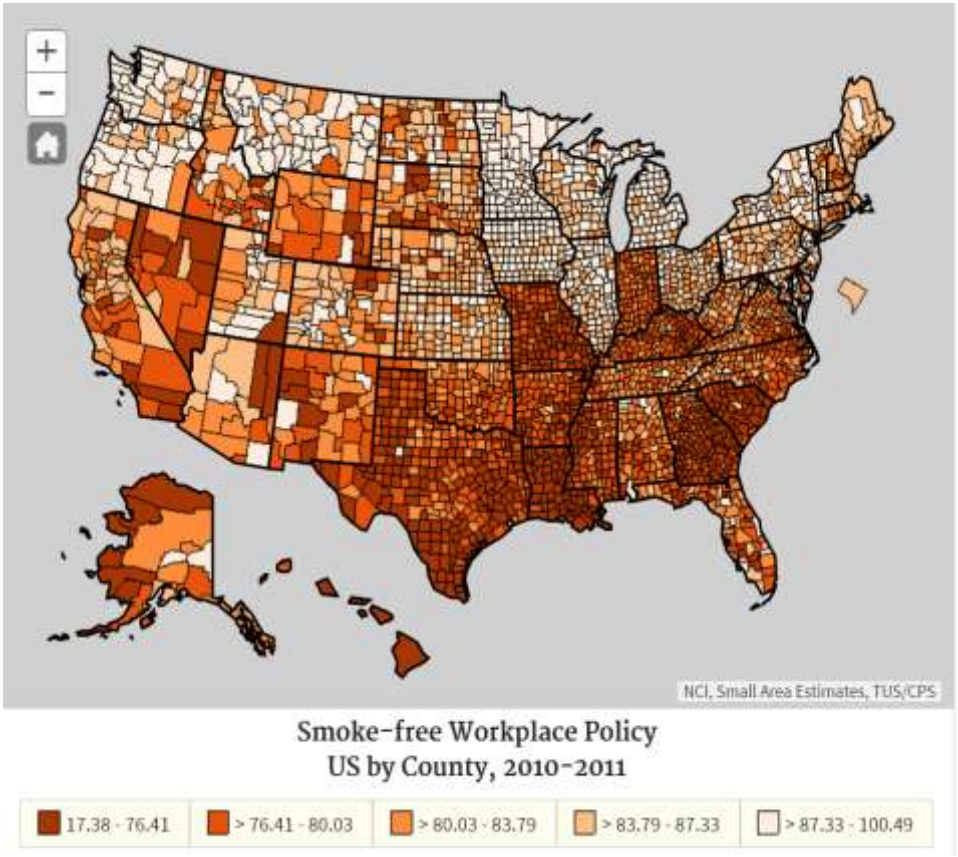
Model-based vs Design Based Estimates for Current Smoking Prevalence –Maryland 2010/11



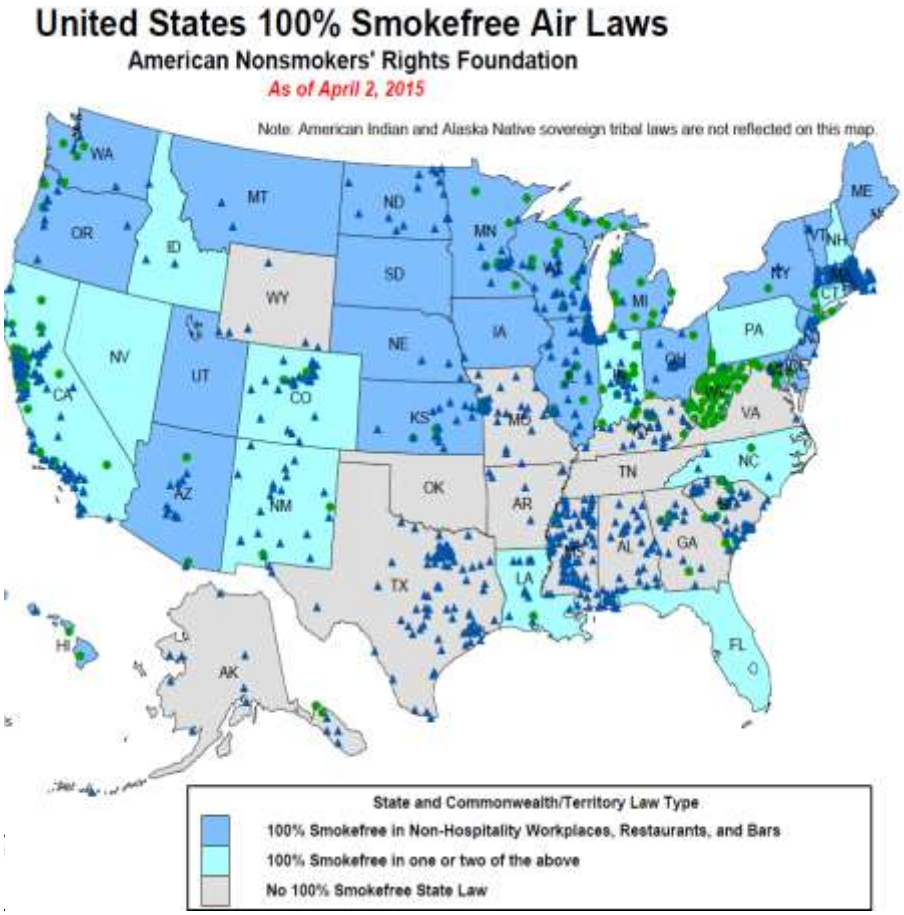


Model-based Estimates for Percent of Population Governed by a Smoke-free Workplace Policy* Among Age 18+: TUS-CPS 10/11

Individual Self-Reported



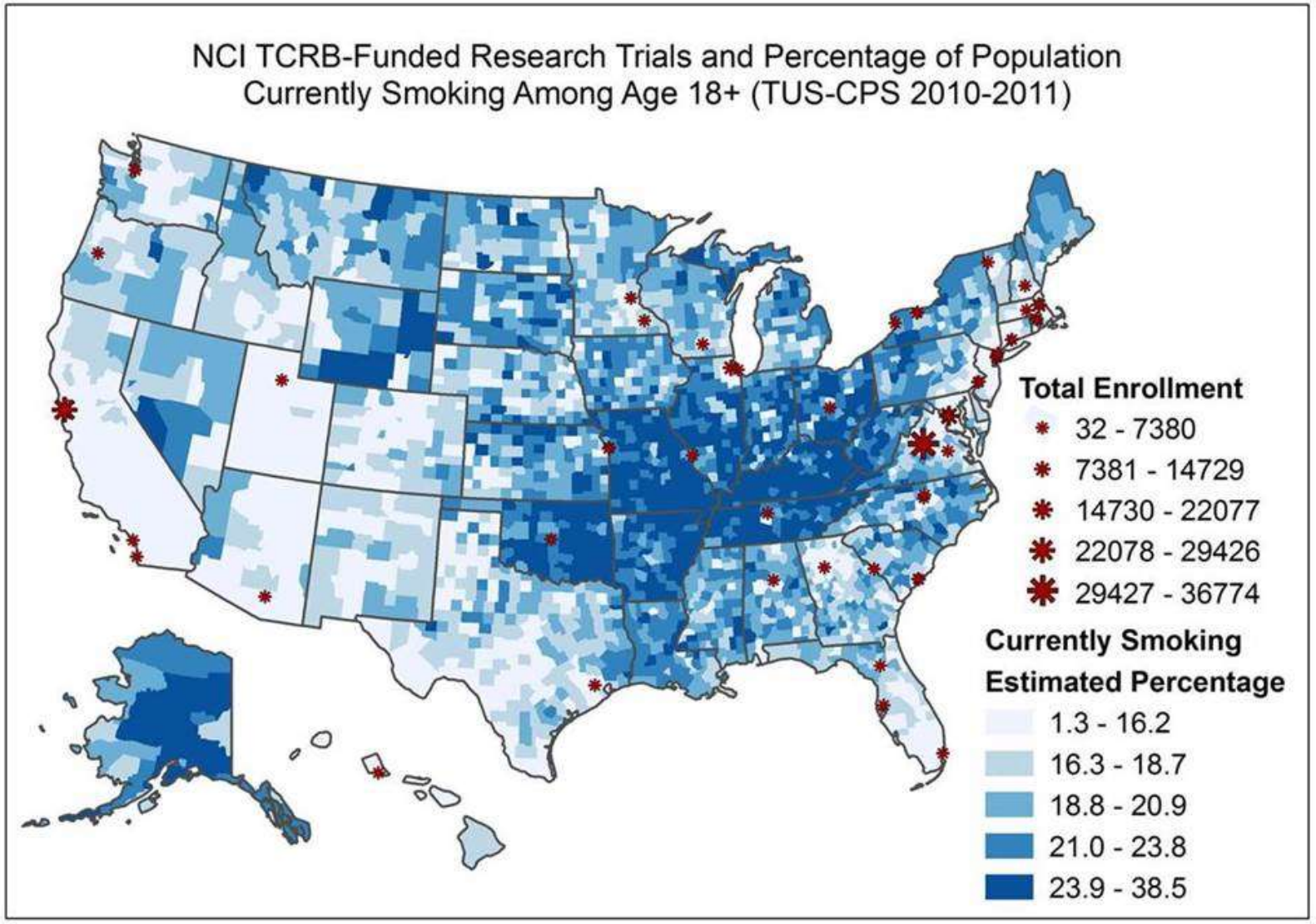
Law Legislations



*Workplace has an official smoking policy: Smoking Not allowed in ANY public areas and work areas

https://sae.cancer.gov/tus-cps/

Applications of the SAE maps



Summary and Discussion

- More details and results are available at https://sae.cancer.gov/tus-cps/.
- The model-based SAE techniques represent an effective means of generating estimates where there is small (or zero) state or county sample.
- The SAE results, which are released and disseminated at several NCI's websites provide a useful resource for the broad cancer surveillance society to fulfill multiple needs.
- We are currently working on estimates for the 2014/2015 data cycle.

Acknowledgement

- Aaron Gilary, Census Bureau
- Partha Lahiri, University of Maryland

Any Questions?

Thank you!

Contact info:

Benmei Liu

liub2@mail.nih.gov