A Practical Framework for Area-level Small Area Estimation

Stephanie Zimmer

All about process

- Audience is statisticians who are not familiar with small-area-estimation (SAE)
- Make SAE approachable and break process into components
- Develop a framework for area-level SAE
- Use concrete examples
  - National Crime Victimization Survey (NCVS)
  - Ohio Medicaid Assessment Survey (OMAS)

Step 1: Understand the requirements

- What are the outcomes?
- What is the small area of interest (domain)? Geographic level, demographic group, etc.
- What is the time period?
- What software can be used? Does it need to be off the shelf? Does it need to be open?
- Anything else?

Step 1: Example

Requirement	OMAS	NCVS
Outcome	Health insurance rates for adults 19+	Victimization rates and prevalence rates for 12 crime types – personal and household
Domain	Ohio counties	State, large counties, large CBSAs
Time period	2017	2009-2015 (3-year estimates for each year)
Software requirement	Off the shelf	Software must be available in the Census RDC
Other		State estimates must sum to national estimates

Step 2: Understand the survey data and sampling design

- Is the design longitudinal, cross-sectional, and/or panel?
- What type of variance estimation is used for direct estimates?
- Are the domain identifiers, sampling variables and all outcomes available on the file?
- Summarize the number of observations per domain and be sure to include domains that may have 0 observations
- Calculate direct survey estimates and standard errors in the domains where possible for later comparison

Step 2: Example - NCVS

- Longitudinal and panel design
- Taylor’s series and replicate weights available
- Geographic identifiers only available on restricted use file
- Direct estimates and standard errors were calculated in a subset of states

Step 3: Identify and obtain potential auxiliary data

- Common auxiliary data sources include:
  - Administrative records
  - Data from large area surveys such as the American Community Survey or censuses
  - Commercial data

- Requirements of auxiliary data
  - Include every domain of interest with matching definitions
  - No missing values

- Obtaining data may take a lot of time including data use agreements

- Use simple regression or decision trees to reduce set of potential variables

Step 3: Auxiliary data example – OMAS potential variables

- ACS 5-year, 2012-2016
  - Adult population (N)
  - Hispanic (p)
  - Am. Indian/Alaskan Native (p)
  - Aged 65+ (p)
  - Non-citizen (p)
  - Housing unit ownership (p)
  - Adults with less than high school (p)
  - Male (p)
- County business patterns
  - Adults employed by non-retail firms (p)
- SAIPE
  - Income (median)
  - Poverty, all ages (p)
  - SNAP receipients (N)
- BEA
  - Per capita income
- BLS
  - Unemployment rate (avg)
- 2010 Census
  - Rural housing units (p)

Step 3: Auxiliary data example – OMAS selected variables

Selection process

- Used linear model with stepwise selection of variables
- AIC used for criteria

Selection results

- BEA
  - Per capita income
- SAIPE
  - Poverty, all ages (p)
- ACS 5-year, 2012-2016
  - Housing unit ownership (p)
  - Adults with less than high school (p)

Step 4: Choose model and assumptions

- Consider the data (longitudinal or cross-sectional)
- Consider the outcome (mean, proportion, other)
- Consider the software limitations and availablity

Step 4: Example - OMAS

- Modified Fay-Herriot hierarchical Bayes model

- Sampling model:
  - Relates the survey direct estimate with the true proportion
  - Includes sampling error component

- Link between parameter of interest and auxiliary variables
  - Linear model using logit transformation
  - Includes random error component

Step 5: Estimation

- Write the code
- Have thorough code review
- Perform model diagnostics
- Check convergence in Bayesian estimation

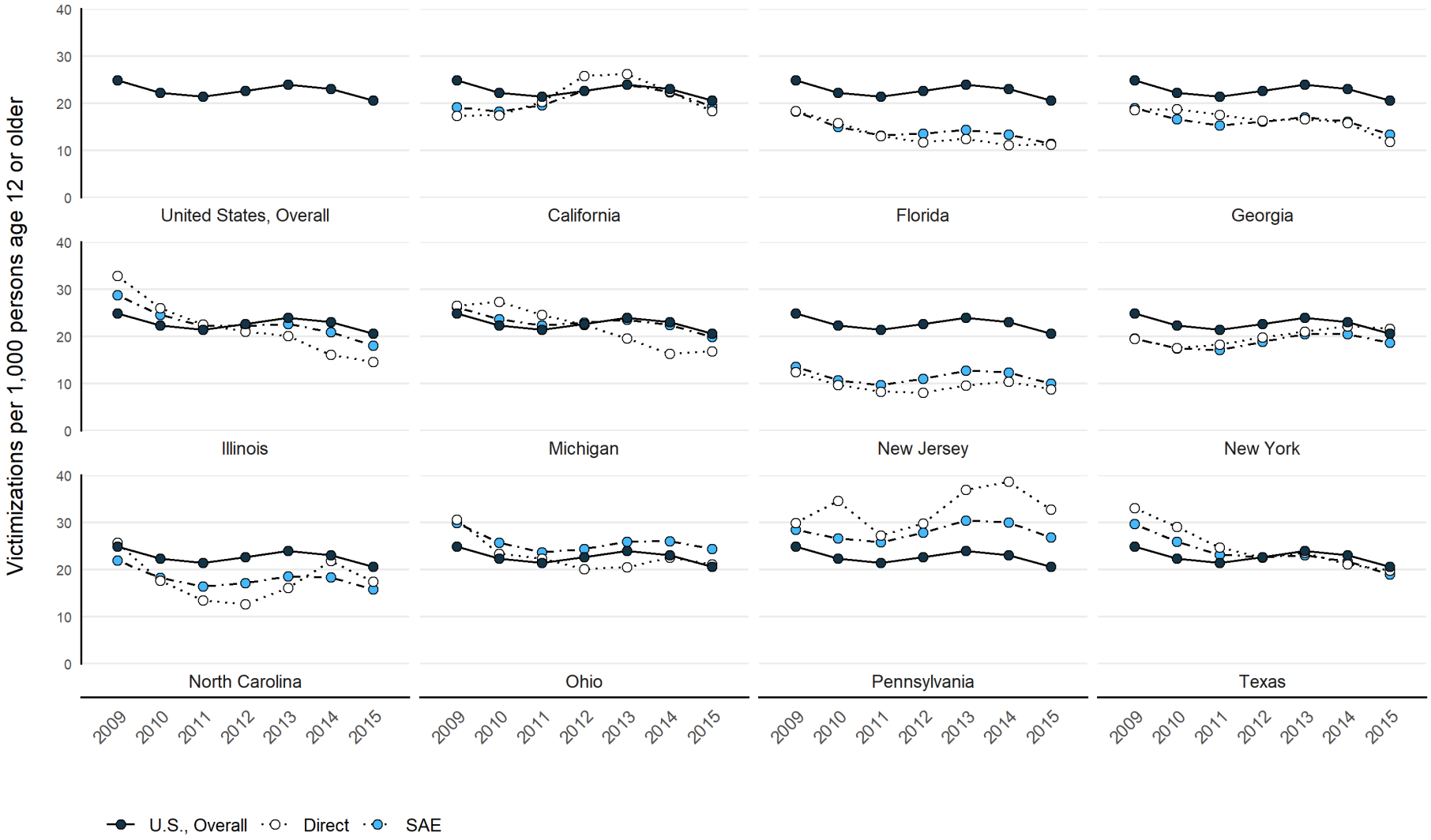
Step 5: Example - OMAS

- R package **R2OpenBUGS** used to estimate parameters which calls **OpenBUGS** from R
- Convergence examined visually with trace plots and with the Gelman-Rubin potential scale reduction factor

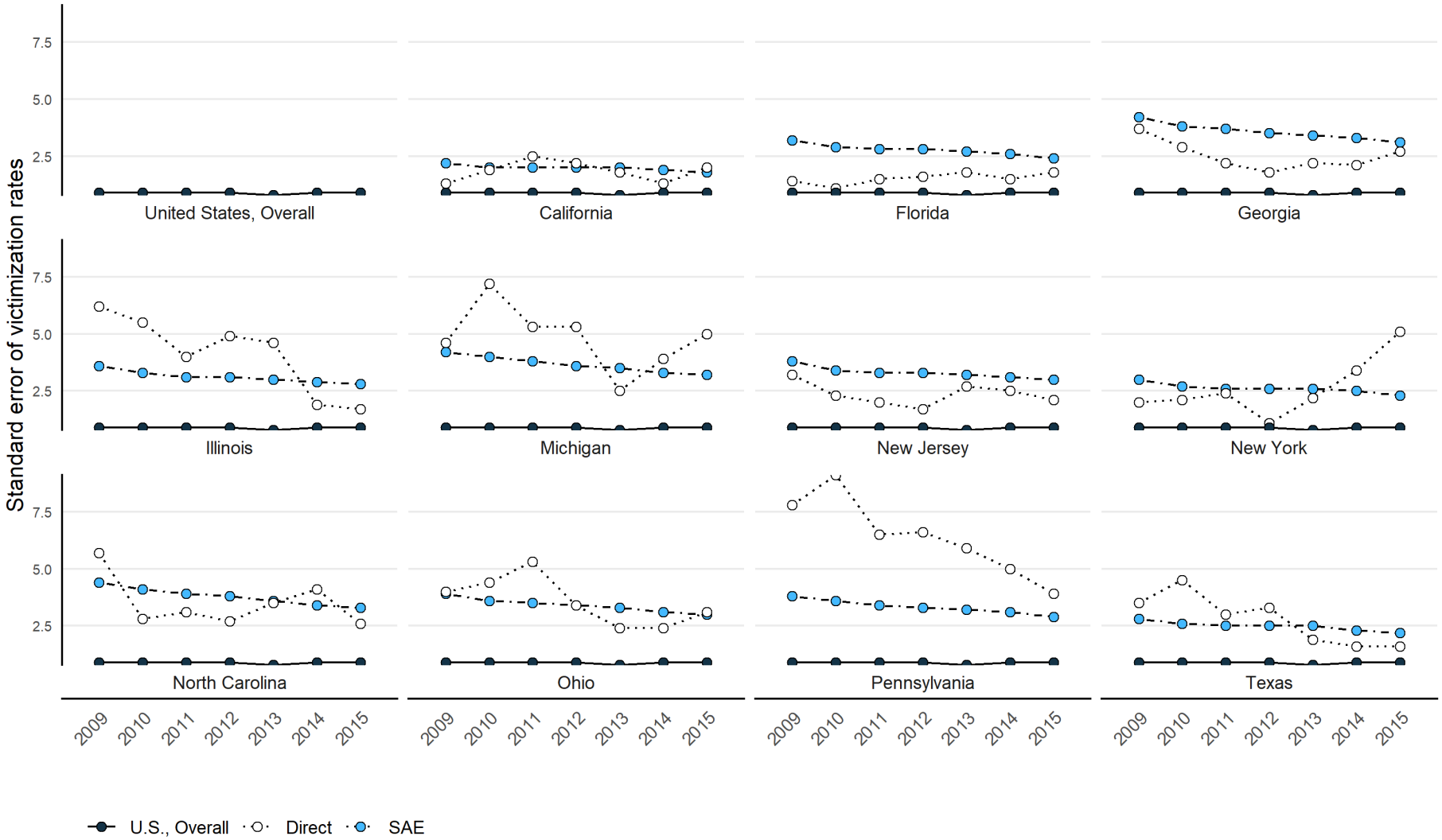
Step 6: Evaluation and validation

- Compare the direct estimates, SAE estimates, and other similar estimates
- Compare precision between direct and SAE estimates

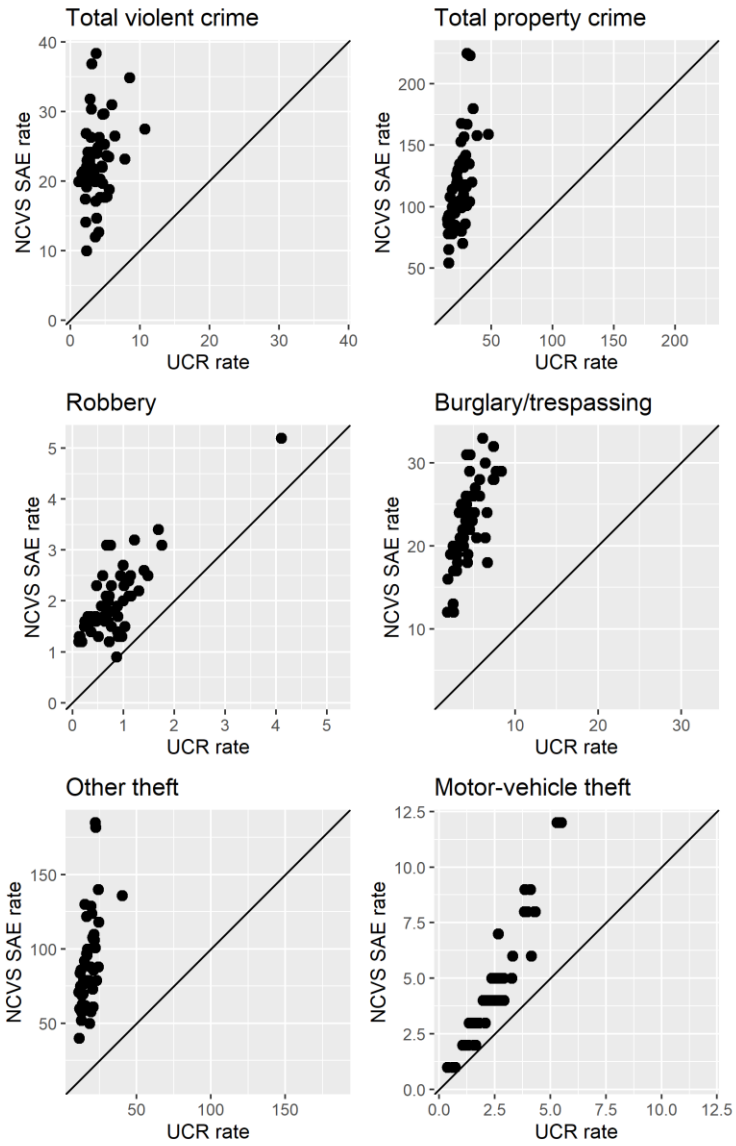
Step 6: NCVS Example: Victimization rate - Direct vs SAE



Step 6: NCVS Example: SE - Direct vs SAE



Step 6: NCVS Example: UCR (Admin data) vs NCVS



Summary

- This paper is geared towards a statistician beginning work in SAE or a non-statistician just wanting to understand the process
- It includes references to seminal works with more details – these are still important
- We want to make this more accessible and open to all

Resources

- Small Area Estimation for the National Crime Victimization Survey: A Guide for Data Processing and Estimation Procedures, 2021

- A Practical Guide to Small Area Estimation, Illustrated Using the Ohio Medicaid Assessment Survey

Thank you

Contact: Stephanie Zimmer| email: sazimmer@rti.org

Dan Liao
Rachel Harter