This report was prepared by RTI International using federal funding provided by the Bureau of Justice Statistics.

Document Title: Small Area Estimation for the National Crime Victimization Survey: A Guide for Data Processing and Estimation Procedures

Authors: Dan Liao, Ph.D., RTI
Stephanie Zimmer, Ph.D., RTI
Marcus Berzofsky, Dr.P.H., RTI

Project Managers: Grace Kena, BJS
Barbara A. Oudekerk, Ph.D., formerly of BJS

Document No.: NCJ 300580

Publication Date: April 2021

Award No.: This project was supported by award number is 2017-MU-MU-K048.

Abstract:

This report was produced by RTI International for BJS under award number 2017-MU-MU-K048. It offers programmers and statisticians guidance on how to use small area estimation (SAE) techniques to develop crime estimates with data from the National Crime Victimization Survey (NCVS) and the Summary Reporting System (SRS) of the FBI's Uniform Crime Reporting (UCR) Program. It also details how to determine the final predictors for deriving the variance-covariance matrix in SAE models and how to use the SAE functions in *R* to compute final subnational estimates for a set of outcomes. Estimates cover victimization and prevalence rates of common crime types. This is a companion report to *Constructing and Disseminating Small Area Estimates from the National Crime Victimization Survey, 2007–2018*.

Disclaimer

The Bureau of Justice Statistics funded this third-party report. It is not a BJS report and does not release official government statistics. The report is released to help inform interested parties of the research or analysis contained within and to encourage discussion. BJS has performed a limited review of the report to ensure the general accuracy of information and adherence to confidentiality and disclosure standards. Any statistics included in this report are not official BJS statistics unless they have been previously published in a BJS report. Any analysis, conclusions, or opinions expressed herein are those of the authors and do not necessarily represent the views, opinions, or policies of the Bureau of Justice Statistics or the U.S. Department of Justice.

This page intentionally left blank.

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

Dan Liao, Ph.D., RTI International
Stephanie Zimmer, Ph.D., RTI International
Marcus Berzofsky, Dr.P.H., RTI International

GRACE KENA AND BARBARA A. OUDEKERK, PROJECT MANAGERS

delivering the promise of science
for global good

RTI International is a trade name of Research Triangle Institute. RTI and the RTI logo are U.S. registered trademarks of Research Triangle Institute.

Acknowledgments

This report was prepared for the Bureau of Justice Statistics by RTI International. In addition to the listed authors, significant contributors include Rachel Harter and Grant Swigart at RTI. Loretta Bohn and Richard Straw copyedited it and Debbie Bond, Judy Cannada, and Loraine Monroe formatted it. The authors thank Dr. Robert Fay, who developed the entire small area estimation methodology and procedures for the National Crime Victimization Survey and graciously shared the materials he developed for this work with us.

CONTENTS

Chapter Page

1. Introduction 8
   1.1 Purpose of This Report 8
   1.2 Background Information on the NCVS and Approaches to Obtain Subnational NCVS Estimates 9
   1.3 SAE for Official Statistics 11
   1.4 SAE for the NCVS 14
   1.5 Organization of the Report 17
   1.6 How to Use This Report and Materials Needed to Implement the SAE Techniques 18
2. Multivariate Dynamic Model 20
   2.1 Introduction 20
   2.2 General Model-Based SAE Approaches 21
   2.3 Rao-Yu Model 21
   2.4 Dynamic Model and Multivariate Dynamic Model 22
   2.5 Using R Package “sae2” to Implement the Dynamic Model 24
      2.5.1 Example 1: Univariate Dynamic Model 24
      2.5.2 Example 2: Multivariate Dynamic Model 26
3. Accommodating the NCVS Sample Design in the Dynamic Model: Estimating Covariance Matrices 29
   3.1 Introduction 29
   3.2 NCVS Sample Design 30
   3.3 Estimating Covariance Matrix for SAE Analysis at the State Level 31
      3.3.1 The Structure of the Covariance Matrix 31
      3.3.2 Accommodating the NCVS Sample Design for SAE Analysis at the State Level 33
      3.3.3 Direct Variance Estimation 33
      3.3.4 Modeling Covariances Over 1997–2015 35
      3.3.5 Modeling Covariances Over 2014–2018 and Beyond 38
      3.3.6 Combining Covariances Across Sets 41
   3.4 Software Implementation 41

4. Use of Auxiliary Information in Developing the SAE Models 43

4.1 Introduction 43

4.2 Auxiliary Information From the UCR SRS 44

4.2.1 Description of the UCR SRS and the NIBRS Data 44

4.2.2 Evaluation of the UCR SRS Data 46

4.3 Regression Prediction of State- and County-Level NCVS Rates From the UCR SRS 48

4.4 Final Selected UCR SRS Predictor for Each Type of NCVS Crime Rate 49

4.5 Suggestions for Selecting Predictors Using NIBRS Data and Other Auxiliary Data Sources for Future NCVS SAE Analyses 50

5. Data Processing Procedures to Prepare the Datasets for Small Area Estimation 52

5.1 Introduction 52

5.2 Preparation of the UCR SRS Data 53

5.2.1 Sources to Download UCR SRS Data 53

5.2.2 Procedure to Prepare Datasets of the UCR SRS State-Level Estimates for the NCVS SAE Modeling Process 55

5.2.3 Procedure to Prepare Datasets of the UCR SRS County-Level Estimates for the NCVS SAE Modeling Process 58

5.3 Preparation of the NCVS Data 62

5.4 Calculating State Controls for the Benchmarking Procedure Using the Census and ACS Data 67

6. Estimation Procedures to Generate the Small Area Estimates 71

6.1 Introduction 71

6.2 The *state_model* Function for State-Level Estimation 72

6.2.1 Arguments of the *state_model* Function 72

6.2.2 Outputs from the s*tate_model* Function 76

6.2.3 Examples of Inputs Using the *state_model* Function 80

6.2.4 Extending to Other Estimates 81

6.3 The *substate_model* Function for Substate-Level Estimation 83

6.4 SAE Program for State-Level Model 86

6.5 SAE Program for Substate-Level Model 89

7. Benchmarking Procedures for State-Level Small Area Estimates 94

7.1 Introduction 94

7.2 Benchmarking Estimates of Crime Subtypes 95

7.2.1 Why Do We Need to Benchmark Estimates of Crime Subtypes? 95

7.2.2 What Crime Subtypes Are Formed and Estimated in the NCVS SAE? 95

7.2.3 Top-Down Benchmarking Approach for Estimation of Victimization Rate 97

7.2.4 Top-Down Benchmarking for Estimation of Prevalence Rate 98

7.3 Benchmarking Victimization Rates and the Estimated Numbers of Victimizations 99

7.3.1 For Violent Crimes 99

7.3.2 For Property Crimes 100

7.4 Benchmarking Prevalence Rates and the Estimated Prevalence Counts 101

7.4.1 For Violent Crimes 101

7.4.2 For Property Crimes 103

7.5 Software Implementation 103

8. Discussion of the State-Level NCVS Small Area Estimation Results 106

8.1 Introduction 106

8.2 State-Level NCVS Small Area Estimates for 2007–2018 107

8.3 Comparisons Between NCVS Small Area Estimates and Direct Estimates for 11 Large States From 2007 to 2015 109

8.4 Comparisons Between the NCVS Small Area Estimates and the UCR Estimated Crime Data 113

8.5 Summary 116

9. References 117

LIST OF TABLES

Table Page

1.1 Examples of Small Area Estimation Projects Conducted by the Federal Agencies in the United States 12

1.2 Use This Report for Different Purposes....19

3.1 Correlation Matrix of the Same Characteristic Within a State 32

3.2 Correlation Matrix of Two Characteristics Within a State 33

4.1 Regression Prediction of County-Level NCVS Violent Crime Rates From the UCR SRS Rates in Self-Representing Counties With the Highest Rates of Complete UCR SRS Reporting, for 1996–2005....48

4.2 UCR SRS Predictor for Each Type of NCVS Crime Rate 50

5.1 An Overview of the Procedure That Prepares Datasets of the UCR SRS State-Level Estimates for the NCVS SAE Modeling Process 55

5.2 An Overview of the Procedure That Prepares Datasets of the UCR SRS County-Level Estimates for the NCVS SAE Modeling Process 58

5.3 An Overview of the Procedure That Prepares Datasets of the NCVS Summary Data for the NCVS SAE Modeling Process....62

5.4 Personal Crime Definitions....66

5.5 Household Crime Definitions 66

5.6 An Overview of the Procedure That Calculates State Controls for the Benchmarking Procedure Using the Census and ACS Data....68

6.1 *state_model* Function Arguments 72

6.2 Key Items of Output From *state_model* Function 76

6.3 *substate_model* Function Arguments....84

8.1 Rates of Total Violent Victimization, by State and Estimation Method, 2007–2015....110

LIST OF FIGURES

Figure Page

3.1 Called-out Functions Under *state_model* Function to Calculate Covariance Matrix for the NCVS SAE Modeling 42

7.1 Dichotomization of Total Violent Crime Into Subtypes..96

7.2 Dichotomization of Total Property Crime Into Subtypes..97

7.3 Process to Obtain the SAE Modeling Results and Perform the Final Benchmarking Procedure...104

8.1 Rates of Violent Victimization, by State and Estimation Method, 2009–2015......112

8.2 SEs and RMSEs of Violent Victimization Rates, by State and Estimation Method, 2009–2015..113

8.3 Comparison of UCR and NCVS SAE 3-Year Victimization Rates, 2016–2018...115

CHAPTER 1. INTRODUCTION

1.1 Purpose of This Report

This report's overarching purpose is to provide thorough instructions for programmers and statisticians on implementing the small area estimation (SAE) approach developed for the National Crime Victimization Survey (NCVS). The approach utilizes a dynamic model described by Fay, Planty, and Diallo (2013) to generate model-based subnational crime estimates for all 50 states, the District of Columbia, and other large geographical areas. The existing NCVS sample has been used along with auxiliary data from the Summary Reporting System (SRS) under the Federal Bureau of Investigation (FBI) through its Uniform Crime Reporting (UCR) Program. Crime estimates include those for crime victimization and prevalence rates of commonly occurring crime types, but these data are not linked to victims' characteristics so as to prevent any disclosure threat.

Special features of this SAE approach include the following:

- use of time-series modeling that employs the NCVS sample data in neighboring years rather than just in the target year to generate overlapping 3-year averages across a 15-year time period;
- accommodation of sampling correlations over time in the model, which can be induced by the NCVS panel design, as well as correlation over time arising from the first-stage selection of NCVS primary sampling units (PSUs);
- use of a restricted maximum likelihood approach to estimate the variance parameters for the model; and
- use of a multivariate model so that crime components and their sum can be modeled jointly to resolve the problem of inconsistency when modeling each component and their sum separately.

This report describes statistical procedures for generating estimates using the SAE models in a manner that is accessible to those less familiar with complex statistical SAE methods. It provides adequate details for an experienced programmer or statistician to process the NCVS data and the UCR SRS data, determine the final predictors to derive the variance-covariance matrix in the SAE models, and use the SAE functions in the R programming language to compute the final subnational estimates for a set of outcomes. It also provides a guide to those interested in using this SAE approach with other national surveys for subnational estimation. Readers are assumed to be knowledgeable enough about statistical modeling and survey statistics to understand the critical technical components of the presented SAE approach. Examples using the R programming language[1] are given throughout this report to illustrate each component of this work for readers who are proficient in statistical programming with R. The more technical and theoretical details of the methodology and the underlying statistical models in this approach were documented by Fay (2021).

Summary of Contents in Chapter 1

What is the purpose of this report?	This report is created to provide thorough instructions for programmers and statisticians on implementing the SAE approach developed for the NCVS. In essence, this is a how-to guide for producing small area estimates of crime victimization and prevalence at the state and substate levels.
Who is the intended audience for this report?	People with substantive expertise in criminology with statistical modeling skills and experienced statisticians who have previously conducted SAE. No experience with SAE is required to understand this report but, for those with experience, chapters will provide detailed information on how to implement the NCVS SAE methodology. Examples in R and supplemental files are provided for readers who are proficient in statistical programming in R to better understand the statistical procedures.
Why is this report being produced?	This report is meant to complement the technical documentation of the NCVS SAE methodology (Fay, 2021) and provide a less technical description of how to operationalize SAE for the NCVS.
How can this report be used?	Depending on the reader's interest, different chapters may be of more use than others. Section 1.6 details which chapters may be of use to readers depending on their methodological or analytical purposes.

1.2 Background Information on the NCVS and Approaches to Obtain Subnational NCVS Estimates

The NCVS is an annual national survey of the civilian, noninstitutionalized population aged 12 or older in the United States. It is the nation's primary source of information on criminal victimization. Survey participants are interviewed on the frequency, characteristics, and consequences of criminal victimization they experienced. They are asked about crimes reported

[1] R is a programming language and open-source environment for statistical computing and graphics. To download and find out more about R, go to https://www.r-project.org/.

and not reported to police, including nonfatal personal crimes (i.e., rape or sexual assault, robbery, aggravated and simple assault, and personal larceny) and household property crimes (i.e., burglary, motor vehicle theft, and other theft).

The NCVS is primarily designed to produce estimates at the national level. The survey uses a stratified, multistage cluster sampling design. PSUs consist of large metropolitan areas, counties, or groups of adjacent counties. Large PSUs are included in the sample with certainty and considered as self-representing because all of them are selected and no PSUs remain in the stratum of large PSUs to represent. The remaining PSUs are considered to be not self-representing. They are stratified into separate strata based on similar geographic and demographic characteristics collected from the decennial census and then randomly selected within each stratum. In 2006 and 2016, the NCVS strata were updated based on the most recent decennial census, the 2000 and 2010 censuses, respectively.

Although the NCVS was originally designed to provide national-level estimates of criminal victimization, the Bureau of Justice Statistics (BJS) recognized an increasing need for victimization data at the state and local levels. Three major reviews of the NCVS program (Biderman, Cantor, Lynch, & Martin, 1986; Groves & Cork, 2008; Penick & Owens, 1976) pointed to the demand that local criminal justice administrators have for empirical information to shape policy. Subnational estimates are of value to both federal and nonfederal data users and stakeholders to examine local variations and trends in both reported and unreported crime to police and to allocate resources for crime victims and crime prevention. Research demonstrated that the NCVS could be enhanced to produce several types of subnational estimates.

Since 2012, BJS has developed multiple approaches for obtaining subnational NCVS estimates, including the following:

- boosting the sample size in large states to obtain direct state-level estimates for certain crime types,
- modeling state-level estimates using existing sample and external sources of data (i.e., SAE), and
- creating generic areas with geocoded identifiers.

For ***approach (1)***, BJS boosted the NCVS core sample in 22 states[2] beginning in 2016 with sample sizes large enough to begin producing 3-year, rolling average, state-level direct estimates of victimization for certain crime types. For ***approach (2)***, SAE has been developed to generate model-based subnational estimates for all 50 states, the District of Columbia, and other large places using the existing NCVS sample along with auxiliary data from the American Community Survey and the FBI's UCR Program. The 1999–2013 state-level small area estimates of crime victimization rates using this approach are presented in Fay and Diallo (2015a). Using essentially the same SAE approach developed for states, these authors also provided 3-year estimates for selected large counties and metropolitan areas. Since then, this approach has been improved to incorporate direct sample boosts starting from 2016 in the NCVS data collection and has been extended to generate estimates for crime prevalence rates. The present report includes these updates to the NCVS SAE modeling work. For ***approach (3)***, Shook-Sa, Lee, and Berzofsky (2015) assessed the coverage and reliability of the NCVS sample in the subnational geographic areas that can be created from the public-use files. Census region, population size, and urbanicity defined these so-called generic areas.

1.3 SAE for Official Statistics

In recent decades, SAE with survey data has become an active research area among federal statistical agencies and other statistical institutes as researchers and policymakers have recognized its potential of informing policy decisions in the absence of sufficient direct sample data. SAE techniques can use statistical modeling and other estimation methodology to "borrow strength" from data across both small areas and time (Ghosh & Rao, 1994). This is done to improve on design-based survey estimates (also known as *direct estimates*) that use only data from sample units in a targeted small area.

To date, a number of federal statistical agencies have conducted a considerable amount of research and development in SAE methods for their own surveys and produced subnational estimates using SAE methods (also see Table 1.1):

[2] The 22 states identified for state-level estimates are Arizona, California, Colorado, Florida, Georgia, Illinois, Indiana, Maryland, Massachusetts, Michigan, Minnesota, Missouri, New Jersey, New York, North Carolina, Ohio, Pennsylvania, Tennessee, Texas, Virginia, Washington, and Wisconsin.

- The U.S. Census Bureau's ***Small Area Income and Poverty Estimates*** program annually produces income and poverty-related estimates for states, counties, and school districts.

- The ***Small Area Health Insurance Estimates*** program (also at the Census Bureau) produces single-year estimates of health insurance coverage status for all counties in the United States by selected economic and demographic characteristics.

- The National Center for Education Statistics produced model-based state and county estimates of adult literacy based on the ***National Assessment of Adult Literacy*** survey.

- The Substance Abuse and Mental Health Services Administration regularly produces state and substate estimates of substance abuse based on the ***National Survey on Drug Use and Health*** using SAE methods.

- The National Center for Health Statistics (NCHS) has conducted several SAE projects, such as generating state estimates of wireless substitution (Blumberg, Luke, Ganesh, Davern, & Boudreaux, 2012).

- NCHS collaborated with the National Cancer Institute, University of Michigan, and University of Pennsylvania to develop methodology that combined estimates from two surveys—the ***Behavioral Risk Factor Surveillance System*** and the ***National Health Interview Survey***—using SAE methods and produced state and county prevalence estimates of cancer risk factors and screening (Liu et al., 2019).

- The Centers for Disease Control and Prevention (CDC), the Robert Wood Johnson Foundation, and the CDC Foundation recently launched the ***500 Cities Project***, which provides city- and census tract-level health indicator estimates using SAE (Kong & Zhang, 2020).

Table 1.1 Examples of Small Area Estimation Projects Conducted by the Federal Agencies in the United States

Agency	SAE Program/Study	Data Year	Produced Small Area Estimates
Census Bureau	Small Area Income and Poverty Estimates (SAIPE) program	1990s to present	Single-year estimates of income and poverty-related estimates for states, counties, and school districts
Census Bureau	Small Area Health Insurance Estimates (SAHIE) Program	1990s to present	Single-year estimates of health insurance coverage status for all U.S. counties, by selected economic and demographic characteristics
National Center for Education Statistics	National Assessment of Adult Literacy (NAAL) survey	1992 and 2003	State and county estimates of adult literacy

(continued)

Table 1.1 Examples of Small Area Estimation Projects Conducted by the Federal Agencies in the United States (continued)

Agency	SAE Program/Study	Data Year	Produced Small Area Estimates
Substance Abuse and Mental Health Services Administration	National Survey on Drug Use and Health (NSDUH)	1999 to present	State and substate estimates of substance use and mental health based on the pooled data from 2 consecutive survey years for state estimates and 3 consecutive survey years for substate estimates
National Center for Health Statistics (NCHS)	National Health Interview Survey (NHIS)	2011	State estimates of wireless substitution
NCHS, National Cancer Institute, University of Michigan, and University of Pennsylvania	Behavioral Risk Factor Surveillance System (BRFSS), NHIS	2008 to 2010	State and county prevalence estimates of cancer risk factors and screening
Robert Wood Johnson Foundation, CDC, and CDC Foundation	500 Cities Project	2016 to present	City- and census tract-level health indicator estimates

CDC = Centers for Disease Control and Prevention; SAE = small area estimation.

Aside from federal statistical agencies in the United States, statistical institutes in other countries as well as international organizations have also produced statistical estimates based on SAE methods. The Office for National Statistics in the United Kingdom has produced small area model-based estimates of income, households in poverty, and house price statistics in small areas within England and Wales. The Italian National Institute of Statistics conducted an SAE study for Italian structural business statistics (Luzi, Solari, & Rocci, 2018). The World Bank collaborated with country teams and developed their PovMap software[3] based on SAE methods for producing poverty maps in more than 20 developing countries.

The advantage of using SAE techniques is that it does not require an additional data collection effort and can produce reliable statistical estimates for a variety of small areas and small subpopulations. However, before allocating resources to produce estimates using SAE techniques, one should carefully take into consideration several caveats:

[3] PovMap software is available at https://www.worldbank.org/en/research/brief/software-for-poverty-mapping.

- Significant time and resources are needed to develop, implement, and evaluate an SAE procedure for a specific statistical project. SAE modeling usually involves a sophisticated and complicated procedure, especially with complex survey data. Tzavidis, Zhang, Luna, Schmid, and Rojas-Perilla (2018) proposed a framework to produce small area official statistics based on three broadly defined stages: (1) specification, (2) analysis and adaptation, and (3) evaluation. Each stage would require technically qualified programmers and experts in SAE to ensure that the SAE techniques are properly specified, analyzed, adapted, and evaluated. As part of the final evaluation, it is sometimes desirable for subject-matter experts to validate the small area estimates by comparing them with direct estimates derived from external data sources. However, validating the small area estimates can present a significant challenge, given that reliable estimates are usually unavailable for the small areas. When comparison estimates exist but are not ideally reliable, making the comparison may still be helpful to see whether the small area estimates are in the same ballpark with the comparison estimates.

- *The effectiveness of the SAE modeling relies heavily on the availability and predictiveness of the auxiliary variables in the SAE models (i.e., how well the auxiliary variables contemporaneously correlate with the outcome variable).* The SAE models need to use auxiliary variables as model predictors. Therefore, the auxiliary data must be available for all small areas, both those contributing observed data (e.g., survey data) in the model and those without any observed data but needing prediction. Even when an auxiliary data source is available for all small areas, the SAE modeling will not be effective and will yield undesirable estimates with poor precision if the auxiliary variables used in the SAE models have no or very poor correlation with the outcome variable. To ensure the success of an SAE activity, one must identify some useful auxiliary data sources with variables that can be highly correlated with the outcome variable(s) first. In addition, if Small area estimates are used for time trend description or analysis, the auxiliary variable may need to reflect the change over time as well.

- *There are data constraints.* Two common types of SAE models are unit-level models and area-level models. Both types require microdata with information at the small area level, the sampling unit level, or both. Data at the sampling unit level usually have restricted access for analysts.

- *Interpreting and communicating the results from an SAE analysis can be a great challenge.* Stakeholders and policymakers may be skeptical about the validity of the small area estimates given its “model-based” feature and feel reluctant to acknowledge the small area estimates as official statistics and use them with confidence.

1.4 SAE for the NCVS

Although the NCVS core sample has been boosted in 22 large states to obtain direct estimates in these states for certain crime types, the NCVS sample data do not guarantee enough sample to facilitate direct estimation in other states, metropolitan areas, or counties or other crime types with sufficient precision. Even for states and areas that have enough sample in a particular year, they may have insufficient sample in other years (e.g., before the sample boosts) due to sampling variation and sample design changes. This insufficiency creates a particular challenge when examining crime trends in a local area using direct estimates based on the NCVS sample data. Therefore, SAE techniques are promising in that they can be used to fill these gaps to produce estimates for all states and some large counties and metropolitan areas without requiring additional data collections in these areas.

The NCVS small area estimates are derived from three basic components: (1) the NCVS sample data collected to support reliable estimates at the national level; (2) the auxiliary information from the UCR SRS data that provides geographically detailed data on reported crimes; and (3) a set of sophisticated SAE models that leverage information from the NCVS data and the UCR SRS data to derive more reliable small area estimates. In view of the caveats regarding the use of SAE techniques as discussed in the previous section, attention to the following six issues is warranted when performing an SAE analysis for subnational estimates of crime victimization and prevalence based on the NCVS.

1. Ensure the predictiveness of the covariates used in the SAE models. The more highly correlated the covariates are with the outcomes to be estimated (e.g., the victimization rate of violent crime), the more effective the SAE models are. The differences in the crime victimization rates between the NCVS and the UCR SRS are well known (Lynch & Addington, 2006; U.S. Department of Justice, 2004). By design, the UCR SRS data reflect only crimes reported to law enforcement agencies (LEAs), but the NCVS data can reveal the “dark figure” of unreported crime (Skogan, 1977). However, Fay and Li (2011) investigated the relationship between the NCVS and UCR SRS rates at the county level, and Li, Diallo, and Fay (2012) later examined this relationship at the state level. Both studies found that the long-term average of the NCVS county or state crime rates for each type of crime was best predicted by a single variable from the UCR SRS (e.g., the UCR SRS robbery crime rate is the best predictor of the NCVS robbery crime rate). Also, in most instances, the NCVS crime rate was best predicted by the corresponding rate from the UCR SRS. Furthermore, the UCR SRS data are available for all states, counties, and metropolitan areas. Therefore, the UCR SRS variables are currently used in the SAE models as covariates to derive the NCVS small area estimates. Chapter 4 in this report summarizes the procedure to select the UCR SRS variables for different outcomes in the NCVS SAE models.

Another emerging auxiliary dataset that might be used to model crime victimization is the FBI's National Incident-Based Reporting System (NIBRS). Unlike the UCR SRS data that contain only aggregated monthly tallies of crimes, the NIBRS data capture details on each single crime incident. In 2018, approximately 44% of U.S. LEAs that participated in the UCR Program submitted their data via NIBRS. The UCR Program is helping other agencies transition to NIBRS with a goal to phase out the traditional SRS data collection and become a NIBRS-only data collection by 2021. The sunset of the UCR SRS data and the emergence of the NIBRS data as a national primary data source on crimes in 2021 will affect the NCVS SAE models. By then, the covariates in the NCVS SAE models can be reselected by statisticians on the basis of the NIBRS data, after summarizing the NIBRS data at the state and substate levels.

2. Avoid estimating rare events. Crimes can be broken up into small crime subtypes, some of which will contain very rare events. Estimating the rate of a rare event is usually challenging and can easily yield estimates with poor precision. When applying the NCVS SAE models to estimate rates for crime subtypes or other outcomes, one should examine the frequency of the outcomes in the entire NCVS sample data to avoid estimating rare events.

3. Clarify that the small area estimates for areas without sample data are synthetic estimates when interpreting the results. The sample design of the NCVS was not stratified by states prior to 2016. Some small population states (e.g., Wyoming) may have little or no sample in a given year in the pre-boost period because the sample design was originally created to support only national estimates. When one is deriving a small area estimate for a small area that has no sample unit at all, the small area estimate is entirely synthetic (i.e., an extrapolation of a model based on other small areas or states). One should be cautious when interpreting these estimates.

4. Align the sum of state-level estimates to the national-level estimates. A benchmarking procedure has been developed for this purpose (see Chapter 7). In this procedure, the national sum of preliminary state-level totals is compared with published NCVS national totals from direct estimation. A raking procedure adjusts the preliminary state-level totals to sum to the published NCVS national totals.

5. Account for the sample design and changes in sample design over time when estimating the variance-covariance matrix. The NCVS sample design is refreshed every 10 years to account for the new decennial census. The variance-covariance matrix in the NCVS SAE models should be re-estimated when the sample design changes. Chapter 3 in this report discusses how to accommodate the NCVS sample design and changes in the design over time when estimating the variance-covariate matrix.

6. Check that time-series data are defined consistently over time. The NCVS SAE models use time-series models, leveraging data across years to improve the reliability of the small area estimates. However, it is possible that the measurements (e.g., survey instruments, definitions of crime types in administrative or survey data) can be modified or even redesigned. The impact of any changes to the data measurements should be carefully assessed and treated in an SAE analysis. For example, in Section 5.2.2, the rape estimates in the UCR SRS data are modified for 2017 and onward because the rape definition in UCR SRS has been revised since 2017.

1.5 Organization of the Report

This report is organized into nine chapters. After this introduction, Chapter 2 provides an overview of the univariate and multivariate dynamic model used for the NCVS SAE models. Chapter 3 describes the history of the NCVS annual sample design and discusses how to estimate the variance-covariance matrix in the SAE models that can accommodate the NCVS sample design. Chapter 4 summarizes the key results from the past work that led to the decision to use the UCR SRS variables as auxiliary data for the NCVS SAE models and the choice of a specific UCR SRS variable for each NCVS crime rate. Chapter 5 describes the data processing procedures to prepare the UCR SRS data, the NCVS data, and the decennial census data for the SAE analysis. Chapter 6 demonstrates how to implement the developed SAE functions to generate small area estimates of crime victimization and prevalence rates at the state and substate levels. Chapter 7 describes the final benchmarking procedures to adjust the state-level small area estimates for consistency with national NCVS estimates. Chapter 8 discusses the NCVS small area estimates and compares them with similar estimates from other sources (e.g., NCVS direct estimates for large states and geographical areas, UCR SRS estimates). References cited throughout the report are listed in Chapter 9.

1.6 How to Use This Report and Materials Needed to Implement the SAE Techniques

This report can be used for various methodological and analytical purposes. When applying the techniques described in this report, one must pay attention to the general caveats of using SAE techniques (Section 1.3) as well as to the special considerations needed when performing an SAE analysis with NCVS data (Section 1.4). Some instructions on the relevance of chapters for different purposes are given in Table 1.2.

Supplemental files that include all the statistical programs mentioned in this report are available upon request for readers who are proficient in statistical programming in R. The supplemental files are created to help readers better understand the statistical procedures and implement the techniques by themselves. In addition to the supplemental files, to produce the NCVS small area estimates, readers need to obtain and process the NCVS data at the Census Bureau or at a Census Research Data Center (RDC) as detailed in Chapter 5.

Finally, please note that this report is intended to complement the technical documentation of the NCVS SAE methodology (Fay, 2021) and serve as a how-to guide for those not fully immersed in the current SAE methods to produce small area estimates of crime victimization and prevalence at the state and substate levels. Readers should refer to (Fay, 2021) for more in-depth information on the SAE methodology behind this work.

Table 1.2 Use This Report for Different Purposes

Purpose	Relevant Chapters and Their Usages
Reproduce the existing NCVS small area estimates of 3-year averages of victimization or prevalence rates, or both, from 2007–2018 at the state level (as shown in Appendix A) or substate level, or both	▪ Chapter 5: Data processing ▪ Chapter 6: Implementing SAE modeling procedure ▪ Chapter 7: Benchmarking the small area estimates
Produce the NCVS small area estimates for the existing outcome variables in a time period that includes 2019 and onward	▪ Chapter 3: Update the covariance matrix based on Section 3.3.5 if the sample design remains the same as the one in 2016–2018; update the covariance matrix based on the entire chapter if the entire sample design is changed ▪ Chapters 5–7: Same as above
Produce the NCVS small area estimates for a new outcome variable at the state or substate level, or both	▪ Chapter 3: Same as above if the time period of interest includes 2019 and onward ▪ Chapter 4: Selecting useful predictors in the auxiliary data to model the new outcome variable ▪ Chapters 5–7: Same as above
Produce the NCVS small area estimates when UCR SRS data are unavailable	▪ Chapter 4: Selecting useful predictors in other auxiliary data (e.g., NIBRS and decennial census data) to model the outcome variables of interest ▪ Chapters 3 and 7: Same as above ▪ Chapter 5: Processing the auxiliary data in a way similar to that for the UCR SRS data ▪ Chapter 6: Using the processed auxiliary data instead of the UCR SRS data in the SAE model functions
Understand or revise the SAE modeling techniques used in the NCVS SAE analysis	▪ Chapter 2: Understand the mathematical formulations of the dynamic models ▪ Chapter 3: Understand how the covariance matrix in the SAE model is estimated ▪ Chapter 7: Understand how the small area estimates of subtypes are benchmarked to agree with the estimate of their aggregated type and how the sum of the state-level estimates agrees with the published national totals

CHAPTER 2. MULTIVARIATE DYNAMIC MODEL

2.1 Introduction

This section briefly summarizes the multivariate dynamic model (Fay et al., 2013) used to derive small area estimates for the National Crime Victimization Survey (NCVS). The dynamic model is a modification of the Rao-Yu model (1992, 1994), described in Section 2.3, that is a small area estimation (SAE) model with special features to handle both time-series and cross-sectional data. The univariate dynamic model and its extended multivariate version are discussed in Section 2.4. The multivariate dynamic approach was chosen to model subtypes of a major class of crime, because of its appealing feature that jointly models the components of crime and their sum (e.g., burglary, motor vehicle theft, and other theft as components of total property crime). This feature resolves the problem of getting inconsistent estimates when using the univariate approach to model components and their sum separately. The R programming language's "sae2" package and its function *eblupDyn*, which was created to implement the dynamic model,[4] are presented in Section 2.5 with two examples, one for the univariate dynamic model and another for the multivariate dynamic model.

Summary of Contents in Chapter 2

	Summary
What is the main point of this chapter?	This chapter describes the SAE modeling approaches used in the NCVS SAE analysis, including both the univariate and multivariate dynamic models. Examples are provided to illustrate these two models.
Why is it important?	The dynamic model is a SAE model with special features to handle both time-series and cross-sectional data. The multivariate dynamic model can be used to jointly model the components of crime and their sum. Theoretical details and practical examples of these models are provided for readers to understand how the small area estimates are modeled in this work.
How it is operationalized?	The R function *eblupDyn* can be used to produce small area estimates of the univariate or multivariate dynamic model.

[4] The abbreviation *eblupDyn* refers to the dynamic model developed by Fay and Diallo (2012). EBLUP stands for "empirical best linear unbiased prediction." For details on the dynamic model, see Section 2.4, and for details on EBLUP, see Rao and Molina (2015).

2.2 General Model-Based SAE Approaches

Two general statistical estimation approaches can be used to produce estimates for outcomes of interest in a local area or a subpopulation, both of which can be considered as a *small area*. These two estimation approaches are (1) *direct estimation* and (2) *indirect estimation*. Direct estimation is the standard design-based estimation approach based on survey data collected directly from the sample units in a small area. The design-based estimation approach sometimes also supplements survey data collected directly with auxiliary or external data sources via some model-assisted approaches (e.g., calibration weighting) to improve the accuracy and efficiency of small area estimates. The direct estimation approach usually can produce reliable results when the number of sample units (i.e., sample size) in a small area is sufficiently large.

However, when the number of sample units is not large enough to guarantee a reliable direct estimate, the second approach (i.e., the *indirect* model-based SAE approach) offers tools to overcome the problem. The main idea underlying the indirect estimation approach is to increase the effective sample size by use of implicit or explicit models that link the related small areas through auxiliary data or from different time periods. The models can also be classified into two major types: (1) area-level models that relate the direct estimates of small areas to the corresponding area-specific auxiliary variables and (2) unit-level models that relate the sampling unit-level information within the small area to unit-specific auxiliary variables, then aggregate the individual unit-level predictions to derive estimates for each small area. Before getting into the mathematical details about the SAE models discussed in this main section, it is worth clarifying that the NCVS models discussed here are for indirect estimation using area-level explicit models with survey data from multiple years of the NCVS and other auxiliary data.

2.3 Rao-Yu Model

Rao (2003) and Fay et al. (2013) summarized the Rao-Yu model (1992,1994), which extended the basic Fay-Herriot model (1979) to handle time-series and cross-sectional data. This model consists of two components—a sampling error model and a linking model. Let y_{it} be the observed mean or total or other sample-based statistic for area *i* and time *t*. The sampling error model can be expressed as follows:

$$y_{it} = \theta_{it} + e_{it}, t = 1, \ldots, T; i = 1, \ldots, D, \tag{2.1}$$

where

θ_{it} is the population value for area *i* and time *t* and
e_{it} is the random sampling error for area *i* and time *t*,
with $\boldsymbol{e_i} = (e_{i1}, \ldots, e_{iT})' \sim N_T(\mathbf{0}, \boldsymbol{\Sigma_i})$.

The linking model that links … can be expressed as follows:

$$\theta_{it} = x'_{it}\beta + v_i + u_{it},$$

where

x'_{it} is a row vector of known auxiliary variables for area *i* and time *t*,
β is a vector of fixed effects for x'_{it},
v_i is a random effect for area *i*,
with $v_i \sim iid\ N(0, \sigma_v^2)$,
u_{it} is a random effect for area *i* and time *t*,
with $u_{it} = \rho u_{i,t-1} + \epsilon_{it}$, $|\rho| < 1$ and $\epsilon_{it} \sim iid\ N(0, \sigma^2)$, and
v_i, ϵ_{it}, and $\boldsymbol{e_i}$ are mutually independent.

Note that the sampling covariance matrix for area i, $\boldsymbol{\Sigma_i}$, need not be diagonal and can accommodate sampling covariances across time within the same area.

2.4 Dynamic Model and Multivariate Dynamic Model

The Rao-Yu model assumes stationarity for the time series,[5] that is, $u_{it} = \rho u_{i,t-1} + \epsilon_{it}$ and thus $|\rho| < 1$, where ρ is the temporal correlation parameter. Fay and Diallo (2012) found this assumption could be questionable when applying the Rao-Yu model for the NCVS state estimates. They modified the Rao-Yu model's linking model and called it the *dynamic model*. Their dynamic model's sampling error model is the same as the one in the Rao-Yu model. The dynamic model's linking model can be expressed as follows:

$$\theta_{it} = x'_{it}\beta + \rho^{t-1}v_i^* + u_{it}^*, \tag{2.2}$$

[5] "Stationarity" means that the statistical properties of a time series (e.g., mean, variance, and autocorrelation structure) do not change over time. This is an important and common assumption in many time-series techniques.

where

v_i^* is a random effect for area *i* at time *t* = 1,
with $\boldsymbol{v_i} \sim iid\ N(0, \sigma_{v*}^2)$,
u_{it}^* is a random effect for area *i* and time *t*,
with $u_{i1}^* = 0$, and
$u_{it}^* = \rho u_{i,t-1}^* + \epsilon_{it}$, for $t > 1$, with $\epsilon_{it} \sim iid\ N(0, \sigma^2)$.

In the dynamic model, ρ is not constrained. When $\sigma_{v*}^2 = \sigma^2/(1-\rho^2)$ and $|\rho| < 1$, the dynamic model is equal to a Rao-Yu model with $\sigma_v^2 = 0$. However, when allowing for $|\rho| > 1$, the dynamic model is more suitable when areas become more disparate over time.

When studying different components of crime and their sum (e.g., burglary, motor vehicle theft, and other theft as components of total property crime), if one applies univariate modeling and derives estimates for each component and its sum separately, then it is very likely that the estimate of the sum does not agree with the sum of the estimates of all the components. To resolve this issue, Fay et al. (2013) implemented a multivariate version of the dynamic model to derive crime estimates with the NCVS data. For the observed sample values of multiple components of crimes, $\boldsymbol{y_{it}} = (y_{it1,} y_{it2,} \ldots)'$ for area *i* and time *t*, the sampling error model can be expressed as follows:

$$\boldsymbol{y_{it}} = \boldsymbol{\theta_{it}} + \boldsymbol{e_{it}},$$

where

$\boldsymbol{\theta_{it}} = (\theta_{it1,} \theta_{it2,} \ldots)'$ is the vector of the population values for area *i* and time *t*, and
$\boldsymbol{e_{it}} = (e_{it1,} e_{it2,} \ldots)'$ is the vector of random sampling errors for area *i* and time *t*.

For *k*th component, the linking model of θ_{itk} can be expressed as follows:

$$\theta_{itk} = x_{itk}' \beta_k + \rho^{t-1} v_{ik}^* + u_{itk}^*.$$

Under the theoretical framework of the multivariate dynamic model, the estimate for the sum of the components modeled is equal to the sum of the estimates for all of the components. This property can guarantee the consistency between the estimate of the sum and the estimates of its components. Therefore, the multivariate dynamic model is used to derive the small area estimates for the NCVS. For more detailed information on this multivariate dynamic model, see Fay et al. (2013).

2.5 Using R Package “sae2” to Implement the Dynamic Model

Fay and Diallo (2015b) developed an R package named “sae2.” One of the key functions in this package, *eblupDyn*, can be used to produce small area estimates of the dynamic model through either a maximum likelihood (ML) or a restricted maximum likelihood (REML) approach, which is the default. This function can fit univariate or multivariate models. In this section, examples from Fay and Diallo (2015b) are used to illustrate how this function can be used to derive small area estimates and what the outputs from this function will provide. Chapter 6 provides more examples on how the *eblupDyn* function can be used with the NCVS data to derive small area estimates at different time points.

2.5.1 Example 1: Univariate Dynamic Model

In this example, a dataset containing 100 observations (from 20 domains and 5 time points) with two variables Y and X is first created. The dataset is sorted in ascending order by time within each domain, which is the sorting requirement for the SAE dynamic model function. Then variable Y is used as the dependent variable, and variable X is used as the independent variable in the SAE dynamic model function, *eblupDyn*. In this function, the number of area domains (D) and time points (T) as well as the sampling covariate matrix (vardir) are also specified. Because the dependent variable has only one variable (i.e., Y), this is a univariate dynamic model.

1. Create a sample dataset with 100 observations (one observation per area and time) for this example.

```
# First, load ‘sae2’ R package.
library(sae2)

# Then set the numbers of areas and time points.
D <- 20 # number of domains (areas)
T <- 5 # number of years (time points)
set.seed(1) # set a seed number for the random process to generate the data.

# Function mvrnormSeries is used to generate data; these data have 100 rows
(D*T=20*5=100) and 2 variables (Y and X), corresponding to equation (2.2) in this section;
the temporal correlation parameter (rho.dyn, $\rho$) is set as 0.9; the v component of the variance
(sigma.v.dyn, $\sigma_{v*}^2$) is set as 1.0; the u component of the variance (sigma.u.dyn, $\sigma^2$) is set as
0.19; and the covariance matrix for the variation due to sampling (sigma.e, $\boldsymbol{\Sigma_i}$ [in equation
(2.1)]) is set as a diagonal matrix, diag(5), which is a 5×5 diagnostic identify matrix.
```

```
data <- data.frame(Y= mvrnormSeries(D=D, T=T, rho.dyn=.9, sigma.v.dyn=1,
                                    sigma.u.dyn=.19, sigma.e=diag(5)), X=rep(1:T, times=D))
```

To take a glance at this generated dataset, print out the first five observations of the dataset corresponding to the five time periods for domain i:

```
> data[1:5,]
           Y X
1 -1.5220543 1
2  0.3983554 2
3  0.1030204 3
4 -1.7400545 4
5 -0.5123742 5
```

2. **Use function *eblupDyn* to implement the dynamic model**

```
result.dyn <- eblupDyn(Y ~ X, D, T, vardir = diag(100), data=data) #vardir defines the
sampling covariance matrix for the direct estimates of the D*T elements (i.e., 100 observations
in this example) of the dependent variable; by default, REML is used.
```

The output from this function, *result.dyn*, includes several components, such as *fit* (containing model-fitting results) and *eblup* (containing the small area estimates, *n* = 100).

Print out result.dyn$fit.

```
> result.dyn$fit
$model
[1] "T: Dynamic, REML"

$convergence
[1] TRUE

$iterations
[1] 17

$estcoef
                  beta  std.error    tvalue    pvalue
(Intercept) 0.24812002 0.28922245 0.8578865 0.3909552
X           0.02908608 0.07087441 0.4103890 0.6815206

$estvarcomp
        estimate std.error
sig2_u 0.0001000 0.1175043
sig2_v 0.6057931 0.3730116
rho    1.0258020 0.1390229

$goodness
          loglike restrictedloglike
        -152.4560         -149.8814
```

Print out the first five observations (five time periods for area *i*) of result.dyn$eblup, which contains the small area estimates for all 100 of the observations in the dataset:

```
> result.dyn$eblup[1:5]
[1] -0.4481181 -0.4377588 -0.4279640 -0.4187163 -0.4098481
```

In this example, the SAE modeling results generated from *eblupDyn* are saved in the object named “**result.dyn**.” The “**result.dyn$fit**” output contains the key model-fitting results.

First, ("T: Dynamic, REML") shows the type of the SAE model used is a (time-series) dynamic model with the REML approach. The model fitted does converge (“$convergence, [1] TRUE”) via 17 iterations. The estimate beta coefficients in equation (2.2), their standard errors, *t*-values, and *p*-values are then given under $estcoef. The estimated random effect coefficients, or ρ, which is the temporal correlation parameter in equation (2.2), and their standard errors are given under $estvarcomp. Finally, some “goodness of fit” statistics are provided in $goodness.
The “**result.dyn$eblup**” contains the final small area estimates for the population values θ_{it} of variable Y for the 100 observations based on the univariate dynamic model results.

2.5.2 Example 2: Multivariate Dynamic Model

In this second example, a dataset containing 100 observations (from 20 domains and 5 time points) with two dependent variables (Y.1 and Y.2) and one independent X variable is first created. The dataset is sorted in ascending order by time within each domain, which is the sorting requirement for the SAE dynamic model function. Then variables Y.1 and Y.2 are used as the dependent variables, and variable X is used as the independent variable in the SAE dynamic model function, *eblupDyn*. In this function, the number of domains (D) and time points (T) as well as the sampling covariate matrix for the direct estimates of the 200 elements (vardir) are also specified. Because there are two dependent variables (Y.1 and Y.2) in the fitted model, the fitted model is a multivariate dynamic model.

1. Create a sample dataset for this example.

```
# First, load ‘sae2’ R package.
library(sae2)
# Then set the numbers of areas and time points.
D <- 20 # number of domains (areas)
T <- 5 # number of years (time points)
set.seed(1) # Set a seed number for the random process to generate the data.

# Similar to Example 1, a dataset is created with two variables (NV=2); the parameters are set
using the same values as what are used in Example 1; and a new parameter (rho.u.dyn) is set
as 0.8 to define the cross-sectional correlation between the two created variables.

data2 <- data.frame(Y= mvrnormSeries(NV=2, D=D, T=T, rho.dyn=.9, sigma.v.dyn=1,
                              sigma.u.dyn=.19, sigma.e=diag(10), rho.u.dyn=0.8), X=rep(1:T,
                              times=D))
```

The first five observations of this generated dataset:

```
1  0.67498592  1.7488437 1
2 -0.03329095  1.0783995 2
3  2.48833063  0.2073316 3
4 -1.20658254 -0.2574515 4
5  0.86241795  0.2585445 5
```

2. **Use function *eblupDyn* to implement the dynamic model.**

```
result.dyn2 <- eblupDyn(list(Y.1 ~ X, Y.2~X), D, T, vardir = diag(200), data=data2) # vardir defines the sampling covariance matrix for the direct estimates of the D*NV*T elements (i.e., 20*3*5=200 in this example) of the dependent variable; by default, REML is used.
```

The output from this function, *result.dyn2*, includes several components, such as fit (containing model-fitting results) and eblup (containing the small area estimates for both Y variables, *n* = 100).

Print out result.dyn2$fit:

```
> result.dyn2$fit
$model
[1] "T: Dynamic, REML"

$convergence
[1] TRUE

$iterations
[1] 21

$estcoef
                        beta  std.error      tvalue    pvalue
(Intercept).1 -0.015176117 0.38041321 -0.03989377 0.9681778
X.1            0.024465377 0.08364524  0.29248979 0.7699122
(Intercept).2  0.219357025 0.37963335  0.57781285 0.5633905
X.2           -0.008053283 0.08710470 -0.09245520 0.9263364

$estvarcomp
          estimate  std.error
sig2_u1 0.07719739 0.10657796
sig2_u2 0.15631376 0.13766760
sig2_v1 1.42747688 0.65474016
sig2_v2 1.38416788 0.64189393
rho     0.81472908 0.07705928
rho_u   0.80417768 0.14705564

$goodness
          loglike restrictedloglike
        -319.8209         -314.5557
```

Print out the first five observations of result.dyn2$eblup, which contains the small area estimates for the population values of both Y.1 and Y.2 for all 100 of the observations in the dataset:

```
> result.dyn2$eblup[1:5,]
        Y.1       Y.2
1 0.7170818 0.9711260
2 0.5906701 0.7746194
3 0.5139054 0.6115564
4 0.3130870 0.3295078
5 0.3296101 0.3308448
```

In Example 2, the SAE modeling results generated from *eblupDyn* are saved in the object named “**result.dyn2**.” The “**result.dyn2$fit**” output contains the key model-fitting results. First, ("T: Dynamic, REML") shows the type of the SAE model used is a (time-series) dynamic model with the REML approach. The model fitted does converge (“$convergence, [1] TRUE”) via 21 iterations. The estimate beta coefficients in equation (2.2), their standard errors, *t*-values, and *p*-values are then given under $estcoef. Because two dependent variables are fitted under the multivariate dynamic model, two sets of beta coefficient results are presented under this element. The first set (“(Intercept)” and X.1) is the estimated beta coefficient results when fitting Y.1 on X, and the second set (“(Intercept)” and X.2) is the estimated beta coefficient results when fitting Y.2 on X. The estimated random effect coefficients, or ρ, which is the temporal correlation parameter in equation (2.2), and the estimated cross-sectional correlation between the two dependent variables are given under $estvarcomp. Finally, some “goodness of fit” statistics are provided in $goodness. The “**result.dyn2$eblup**” contains the final small area estimates for the population values of both variables Y.1 and Y.2 of the 100 observations based on the multivariate dynamic model results. The small area estimates for the population values of the sum of Y.1 and Y.2 are the sum of the small area estimates for both values.

CHAPTER 3. ACCOMMODATING THE NCVS SAMPLE DESIGN IN THE DYNAMIC MODEL: ESTIMATING COVARIANCE MATRICES

3.1 Introduction

The target population of the National Criminal Victimization Survey (NCVS) is U.S. residents age 12 or older residing in housing units (HUs) or group quarters (GQs) such as dormitories, rooming houses, and religious group dwellings (Bureau of Justice Statistics [BJS], 2017). The NCVS estimates are correlated over time because of the special features of the NCVS sample. The multivariate dynamic modeling method used for the NCVS small area estimation (SAE) analysis, as described in Chapter 2, also models multiple outcome variables that can be correlated with each other. This chapter details how these correlations can be accommodated in the covariance matrix under the dynamic models. It is worth noting that the SAE procedures (see Chapter 6) employ the NCVS sample data in neighboring years (e.g., 2014–2016) rather than just in the target year (e.g., 2015) to generate overlapping 3-year averages as small area estimates across a 15-year time period, but the covariance matrix is estimated based only on the NCVS sample data in the target year.

In this chapter, Section 3.2 discusses the special features of the NCVS sample design. Section 3.3 illustrates the model assumptions and the resulting structure of the covariance matrix for the state-level estimation. Section 3.4 provides instructions on software implementation to obtain the estimated covariance matrices via the developed R program files (provided among the supplemental files for this report). The estimation procedure of the covariance matrices for the county-level estimation is similar to the procedure for the state-level estimation. The estimated covariance matrices for the county-level estimation can also be used for the Core-Based Statistical Area (CBSA)-level estimation. R program files developed to obtain the estimated covariance matrices for the county-level estimation are also discussed in Section 3.4 and provided as supplemental files.

Summary of Contents in Chapter 3

	Summary
What is the main point of this chapter?	This chapter describes how the covariance matrices in the NCVS SAE models are estimated.
Why is it important?	The NCVS estimates are correlated over time because of the NCVS's rotating panel sample design. The multiple outcomes in the same SAE model are also correlated with each other. The covariance matrix in the SAE model needs to account for these correlations to achieve accurate estimation results.
How is it operationalized?	The SAE modeling function, *state_model* (for state-level estimation) or *substate_model* (for substate-level estimation), calls several other functions to estimate the covariance matrix given the model and years specified within the modeling function.

3.2 NCVS Sample Design

The NCVS sample design is a two-stage sample. The first stage selects a sample of primary sampling units (PSUs). In the NCVS, a PSU is either a large metropolitan area, county, or group of bordering counties. The first-stage sampling occurs once every 10 years. The 2000 census design, which sampled PSUs using population data from the 2000 Census, was implemented for interviews from 2006 through 2015. The 2010 Census design, which sampled PSUs using population data from the 2010 Census, was used for interviews starting in 2016. Therefore, the NCVS data have the same first-stage sample PSUs from 2006 to 2015, but some first-stage sample PSUs in the 2016 NCVS are different from the ones in 2015.

Within the first-stage sample PSUs, some PSUs that are large, and thus important contributors to estimates, were selected with certainty. They are known as self-representing (SR) PSUs because they represent themselves and no other PSUs. In addition, all PSUs within a large CBSA are SR PSUs regardless of their size. Other PSUs with smaller populations were selected through a stratified probability-proportional-to-size (PPS) sampling scheme, where the measure of size was total population from the decennial census. The selected smaller PSUs represent themselves and all the other PSUs in the same stratum. They are referred to as non-self-representing (NSR) PSUs.

The second-stage sampling selects HUs every year and GQs every 3 years within the selected first-stage PSUs. The NCVS survey adopts a rotating panel design in which persons in each HU or GQ are interviewed every 6 months for 3 successive years, for a total of seven interviews. Sampling units (HUs or GQs) that have completed the seven interviews are replaced by new sampling units that rotate into the sample. If a household moves out of a sampling unit during the 3-year interview period, the incoming household replaces it in the survey. The sampling unit is the address rather than the persons in the household at the time of the first interview.

These special features of the NCVS sample design affect the random sampling error terms, denoted as $\boldsymbol{e_{it}}$ in Chapter 2 in the dynamic models that are used to derive small area point estimates and their estimated variances. The sampling covariance matrix is not diagonal and will need to accommodate sampling correlations among the estimates over time due to (a) the first-stage PSU sampling, (b) the panel design, and (c) the method of replacing retired sampling units.

3.3 Estimating Covariance Matrix for SAE Analysis at the State Level

3.3.1 The Structure of the Covariance Matrix

As illustrated in Chapter 2, when modeling a single characteristic (e.g., crime type) in the dynamic models, one assumes that the random sampling error e_{it} for area *i* and time *t* follows a normal distribution that can be expressed as

$$\boldsymbol{e_i} = (e_{i1}, \dots, e_{iT})' \sim N_T(\mathbf{0}, \boldsymbol{\Sigma_i}), t = 1, \dots, T; i = 1, \dots, D, \tag{3.1}$$

where $\boldsymbol{\Sigma_i}$ is the covariance matrix for domain or area *i*. The correlation of sampling errors between any two different areas is assumed to be zero in the dynamic models. Therefore, the covariance matrix for the entire model is a block diagonal matrix that can be expressed as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma_1} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \boldsymbol{\Sigma_D} \end{pmatrix} \tag{3.2}$$

and $\boldsymbol{e} = (e_{11}, \dots, e_{1T}, \dots, e_{D1}, \dots, e_{DT})' \sim N_T(\mathbf{0}, \boldsymbol{\Sigma})$.

When multiple *M* characteristics (e.g., multiple components of crimes), are being modeled, the random sampling error can be expanded to be expressed as $\boldsymbol{e_{it}} = (e_{it1}, \dots, e_{itM})'$ for area *i* and time *t* with $\boldsymbol{e_i} = (\boldsymbol{e}_{i1}, \dots, \boldsymbol{e}_{iT})' \sim N_T(\mathbf{0}, \boldsymbol{\Sigma_i})$. For the state-level estimation, the elements on the main diagonal of $\boldsymbol{\Sigma_i}$ are the *variances* of estimates for the same characteristic in the same year (time) and the same state (area). The off-diagonal elements are the *covariances* between estimates of the same characteristic in different years or estimates between different characteristics within the same state. For simplicity in this chapter, all the elements in the covariance matrix, including variances, are referred to as *covariances*, unless a distinction is necessary.

Given the complex structure of the covariance matrix, three main assumptions are made for simplicity when constructing the covariance matrix under the dynamic models for the state-level estimation:

1. The correlation of sampling errors between any two different areas is assumed to be zero in the dynamic models as illustrated above.

2. The correlation of sampling errors between 2 years for the same characteristic within a state is the same regardless of those years. For example, the correlation between 2007 and 2008 is assumed to be the same as the correlation between 2008 and 2009. This assumption leads to some simplification and allows estimated correlations to be averaged across many observations. Nevertheless, because of the substantial complexity of the 2000 census redesign that was first implemented in 2005 and 2006, correlations between 1999–2004 and 2005–2013 (and onward) are set to zero. Thus, a correlation matrix of the same characteristic within a state has the structure shown in Table 3.1.

Table 3.1 Correlation Matrix of the Same Characteristic Within a State

Year	2003	2004	2005	2006	2007	2008	2009
2003	1	C_1	0	0	0	0	0
2004	C_1	1	0[a]	0	0	0	0
2005	0	0[a]	1	C_1	C_2	C_3	C_4
2006	0	0	C_1	1	C_1	C_2	C_3
2007	0	0	C_2	C_1	1	C_1	C_2
2008	0	0	C_3	C_2	C_1	1	C_1
2009	0	0	C_4	C_3	C_2	C_1	1

Notes: C_1 represents a correlation of estimates with a 1-year lag, C_2 represents a correlation of estimates with a 2-year lag, and so on.
a: Because of the substantial complexity of the 2000 census redesign that first implemented in 2005 and 2006, correlations between 1999–2004 and 2005–2013 are set to zero.

3. The correlation of sampling errors between any two different characteristics within a state is estimated assuming that the correlation between the two characteristics at a given lag of years is constant regardless of the pair of years. For example, the correlation between Characteristic A in 2007 and Characteristic B in 2009 is assumed to be the same as the correlation between Characteristic A in 2010 and Characteristic B in 2012. Table 3.2 shows the correlation matrix of two characteristics within a state.

Table 3.2 Correlation Matrix of Two Characteristics Within a State

Year	2007(y1)	2007(y2)	2008(y1)	2008(y2)	2009(y1)	2009(y2)
2007(y1)	1	R_{12-0}	C_{1-1}	R_{12-1}	C_{1-2}	R_{12-2}
2007(y2)	R_{12-0}	1	R_{12-1}	C_{2-1}	R_{12-2}	C_{2-2}
2008(y1)	C_{1-1}	R_{12-1}	1	R_{12-0}	C_{1-1}	R_{12-1}
2008(y2)	R_{12-1}	C_{2-1}	R_{12-0}	1	R_{12-1}	C_{2-1}
2009(y1)	C_{1-2}	R_{12-2}	C_{2-1}	R_{12-1}	1	R_{12-0}
2009(y2)	R_{12-2}	C_{2-2}	R_{12-1}	C_{2-1}	R_{12-0}	1

Notes:
The variables y1 and y2 stand for two different characteristics that are included as outcome variables in the same multivariate model.
C_{1-1} represents a correlation of characteristic y1 between 2 years with a 1-year lag, and C_{1-2} represents a correlation of characteristic y1 with a 2-year lag.
C_{2-1} represents a correlation of characteristic y2 between 2 years with a 1-year lag, and C_{2-2} represents a correlation of characteristic y2 with a 2-year lag.
R_{12-0} represents a correlation between characteristics y1 and y2 within the same year, R_{12-1} represents a correlation between characteristics y1 and y2 from 2 different years with a 1-year lag, and R_{12-2} represents a correlation between characteristics y1 and y2 from 2 different years with a 2-year lag.

The following subsections will discuss how the correlations between 2 consecutive years for the same characteristic and the correlations between two different characteristics within a state are modeled and how the components in the covariance matrix are calculated to construct the covariance matrix for the dynamic models.

3.3.2 Accommodating the NCVS Sample Design for SAE Analysis at the State Level

Because the NCVS sample design was not stratified by state until 2016, some states may not have adequate sample sizes in a certain year to produce stable direct variance estimates. Therefore, the modeling of the covariance matrix starts at a higher geographical level to obtain direct covariance estimates for SR and NSR areas first. Then, the results are averaged and distributed down to the state level. Section 3.3.3 illustrates the calculations for direct variance estimation, including adjusting the design variables in the NCVS data from different years. Section 3.3.4 describes the covariance modeling from 1997 to 2015. Section 3.3.5 addresses the sample redesign implemented in 2016 for the 2010 Census and provides details on the covariance modeling from 2014 to 2018. Section 3.3.6 illustrates how to combine estimated covariance matrices derived for different ranges of years.

3.3.3 Direct Variance Estimation

There are two methods for direct estimation of variances in the NCVS: Taylor's series linearization (TSL) and a replication method using replicate weights. These two methods, on average, lead to similar estimates of variance for national estimates. The TSL method was chosen over the replication method for three major reasons:

1. Replicate weights are not available for all years; they were first calculated in 2011.
2. Replicate weights create larger data files and require more computation time to calculate variances than TSL. Therefore, TSL variance estimation is faster.
3. There was no coordination between the replicate weights for 2015 and the replicate weights designed for 2016 and after (i.e., across the phase-in period of new Census PSUs[6]), which would require a specialized solution to derive covariances between the two time periods. Williams et al. (2015) discussed how replicate weights would need to be created to estimate covariances across different designs; that approach is not currently implemented in NCVS.

The NCVS data files maintained by the Census Bureau, and available through restricted-use agreements, have all the necessary design variables[7] for TSL variance estimation: the pseudo-stratum and the half-sample code (also referred to as SECUCODE). Some variables need to be retained or modified to capture the covariance between the estimates up to 2015 and those after 2016. These actions include the following:

- **Retaining the 2015 pseudo-stratum and half-sample code for the 2015 data kept in the 2016 revised analysis file:** Because households (including GQs) are interviewed every 6 months for seven interviews, some sampling units in the sample in 2015 could remain in the sample until as late as 2018. The 2016 revised analysis file includes some data from 2015 to bridge the 2000 and 2010 designs to adjust for a possible impact of the large number of first-time interviews (see Morgan & Kena, 2018, pp. 3–4, for more details). The 2016 revised analysis file is used both for official national estimates and as input to the SAE modeling. For variance estimation inputs into the SAE algorithm, the 2015 data included in the 2016 revised analysis file retained their 2015 pseudo-stratum and half-sample codes, as did any sample in SR areas in 2016 and 2017. This was done to preserve the covariance between the units that were being phased out from the prior design.
- **Modifying the pseudo-strata so that they do not overlap between the 2000 and 2010 designs:** To enable production of state-level direct estimates based on 3-year rolling averages, BJS boosted the NCVS core sample in 22 states. Starting with 2016, a value equal to 1,000 times the FIPS state code is added to the original pseudo-stratum code for the supplemented states in the data files, thus strictly reflecting the state stratification in the variance calculation. For the 2010-based sample design in the remaining states, 500 is added to the original pseudo-stratum code in the data files, which is sufficient to distinguish these cases from any in 2015 or earlier years.

Once the codes for pseudo-stratum and half sample are set up, the direct variance and covariance estimates for different characteristics can be calculated using standard survey

[6] The "phase-in period" refers to the sixth year of each decade (e.g., 2016) when new PSUs selected based on the latest decennial census are rotated into the NCVS sample design. For example, PSUs based on the 2010 Census were phased in in 2016.

[7] The public use files for 2016 do not include the design variables necessary for TSL estimation.

software such as the *survey* package in R (Lumley, 2020). The survey package functions can calculate the TSL variance estimates of victimization rates and prevalence rates, both of which are ratios of estimated totals. The victimization rate is the ratio of the estimated number of victimizations to estimated population size, and the prevalence rate is the ratio of estimated number of victims to the estimated population size. Although the numerator and denominator in some of the estimated ratios use different weights, they share the same sample design. The standard linearization variance estimation approach for ratios can be used to produce direct variance-covariance estimates for both victimization rates and prevalence rates.

Standard software, such as the survey package in R, can give direct variance and covariance estimates at higher geographical levels such as SR or NSR areas, but these estimates are not sufficient for small areas because there may be very small sample sizes for some states in certain years, especially for characteristics with rare events. Hence, covariances (including variances) are modeled rather than calculated directly. The method for modeling the covariances for 1997–2015 differs from that for estimating covariances from 2014 and later, so these methods are discussed separately in the following sections. There is overlap in this period because if an estimate is requested for 2014–2018, it has units under both designs and is treated differently from an estimate based on years solely in 2015 and earlier. The four sets of years for variance modeling are 1997–2004, 2005–2015, 2014–2018, and 2016 and beyond.

3.3.4 Modeling Covariances Over 1997–2015

As discussed in Section 3.3.1, the covariance estimates within a state between different time points and between different characteristics are nonzero. However, the covariances between 1997–2004 and 2005–2013 are set to zero because of the substantial complexity of the 2000 census redesign in 2005. Thus, when the covariance estimates are ordered by state, the corresponding covariance matrix is assumed to be block diagonal. Three modeling steps are conducted to model the covariance matrix for each state:

Step 1: The directly estimated SR and NSR covariance matrices are converted to correlation matrices, which are then separately modeled and smoothed.

Step 2: The smoothed correlation matrices are then converted into smoothed covariance matrices based on the directly estimated SR and NSR variances for each year.

Step 3: For any given state, its modeled estimate of the covariance is a combination of the SR and NSR smoothed covariances and the state's expected sample sizes in SR and NSR areas.

These steps are detailed as follows.

Step 1—Estimating smoothed correlation matrices for SR and NSR estimates separately

For the covariance matrix of each SR or NSR, the sampling variances along the diagonal of the matrix are calculated directly from the direct variance estimation as described in Section 3.3.3 to reflect the sample size and design in the corresponding year. Because the NCVS sample size varies across years, the sampling variances are not modeled. However, estimating the covariances (i.e., the off-diagonal elements in the variance-covariance matrix) is not so straightforward. The correlation of the corresponding pair of a covariance needs to be estimated first, as illustrated in this step.

The correlation of a pair of years is estimated as the average of correlations between any 2 years that have the same difference in years as for this pair. For example, the correlation between a single-year difference in 1999–2013, denoted as C_1 in Table 3.1, is estimated as the average correlation for the pairs of years: 1999/2000, 2000/2001, 2001/2002, 2002/2003, 2003/2004, 2006/2007, 2007/2008, 2009/2010, 2010/2011, 2011/2012, and 2012/2013.[8] For a 2-year difference in 1999–2003, denoted as C_2 in Table 3.1, the correlation is estimated as the average of the correlations of the following pairs: 1999/2001, 2000/2002, 2001/2003, 2002/2004, 2006/2008, 2007/2009, 2009/2011, 2010/2012, and 2011/2013. This continues for correlations of 3-year, 4-year, and up to 7-year differences to estimate average correlations. The initial correlation matrix for 2003–2009, for example, can be written as the one in Table 3.1 for a given state and characteristic.

After average correlations are obtained for each difference in years, they are further smoothed by fitting a linear regression to them. The dependent variable in the regression model is the average correlation; the independent variable is the difference in years. The regression

[8] The correlations between years 1997–2004 and 2005–2018 (and onward) are set as 0 because of the complexity of the 2000 census redesign that was first implemented in 2005 and 2006. Year 2005 is also excluded from the averaging because its design information on the internal file is not consistent with that of 2006 and subsequent years.

model is weighted by the number of paired years used in the calculation of each difference. In other words, differences of 1 year are given the most weight, because they are supported by the most data, and the largest differences with only one or a few observations are given less weight. The predicted values from the regression are then assumed to be the true correlation, but negative predictions are set to zero.

When two characteristics are being modeled, the covariance between them in the same year is kept fixed and estimated via direct variance estimation, whereas correlations between different years are modeled in a way similar to the one for a single characteristic. The correlations are assumed symmetric, so the correlation between one characteristic at 1 year and another characteristic at a second year is assumed to be the same as the correlation of the other characteristic at the first year and the first characteristic at the second. Correlations between different characteristics in the same year are preserved unadjusted. This process is done for SR and NSR areas separately to produce two correlation matrices.

Step 2—Smoothed correlation matrices converted to covariance matrices

After the correlations have been modeled, they are converted into modeled covariance matrices based on the directly estimated variances. Consequently, the directly estimated covariances between different characteristics in the same year are preserved in multivariate applications. Thus, the modeled covariances are a mixture of directly estimated elements in the same year and smoothed elements between different years. This process is done for SR and NSR areas separately to produce two covariance matrices, namely, C_{SR} and C_{NSR}.

Step 3—Combining estimated SR and NSR covariance matrices

Within each state, the SR and NSR covariance matrices are combined. They are combined using weights derived from their *expected* sample sizes in 1997–2015 because the sample size in the combined NSR areas of a state was random and highly variable across years. At the extreme, some small states might not include any NSR PSUs. For a given state, i, let $n_{SR,i}$ and $n_{NSR,i}$ denote the expected sample size in the state, and let n_{SR} and n_{NSR} represent the actual SR and NSR sample sizes. Letting

$$f_i = \frac{n_{SR,i}}{n_{SR,i}+n_{NSR,i}}, \tag{3.3}$$

the modeled state-level covariance is

$$C_i = \frac{n_{SR}}{n_{SR,i}} f_i^2 C_{SR} + \frac{n_{NSR}}{n_{NSR,i}} (1 - f_i)^2 C_{NSR}. \tag{3.4}$$

The methodology to estimate the expected sample sizes in the SR and NSR areas is provided by Fay and Li (2012). The model for expected sample sizes is based on publicly available information and reasoned guesses of the design. Additionally, a file *modeled_design.csv* is included in the supplemental files to use in code.

3.3.5 Modeling Covariances Over 2014–2018 and Beyond

3.3.5.1 State Supplementation in the 22 Largest States Since 2016

In 2016, the NCVS design introduced state supplementation in the 22 largest states, which increased sample size in these states.[9] One of the results of the supplementation was a design that was no longer approximately self-weighting. Additionally, this design change necessitated a change in the covariance modeling. To mitigate the variation in weights starting in 2016, states were divided into two groups:

State Group 1: 18 of the 22 supplemented states—all but California, Florida, New York, and Texas.

State Group 2: California, Florida, New York, Texas, the non-supplemented states, and the District of Columbia.

Because the average weights in Group 1 were smaller than those in Group 2, this division creates two groups with relatively homogeneous weights within each group but different weights between the groups. Hence, the covariance matrices are modeled differently by different groups.

3.3.5.2 Overview of the Entire System to Calculate Covariance Matrices

To better illustrate how the covariances are modeled for years in 2014–2018 and even beyond, the entire system to calculate the covariance matrix is presented as a whole, so that one can understand how the covariance estimation method for years in 2014–2018 is different from the method for prior years, as described in Section 3.3.4. In summary, the current system

[9] The 22 states with supplemented sample sizes were Arizona, California, Colorado, Florida, Georgia, Illinois, Indiana, Maryland, Massachusetts, Michigan, Minnesota, Missouri, New Jersey, New York, North Carolina, Ohio, Pennsylvania, Tennessee, Texas, Virginia, Washington, and Wisconsin.

calculates four different sets of direct covariance matrices depending on the range of years specified for the SAE model, given the change of sample design over time.

Set 1. SR and NSR covariances for years in 1997–2004

- **Method to estimate covariances:** Covariances for this period are modeled separately for SR and NSR units (Section 3.3.4).
- **Pairs of years for which the average of correlations are estimated and modeled:** As described in Step 1 in Section 3.3.4, correlations between any 2 years in 1997–2004 are estimated and modeled. Then, the modeled (smoothed) correlation matrices are converted to covariance matrices based on the directly estimated variances.
- **Creating state-level covariance:** For any given state, the SR and NSR covariances are combined using the estimated sample size for the state (i.e., a weighted average based on the estimated sample size in the NR and NSR groups).

Set 2. SR and NSR covariances for years in 2005–2015

- **Method to estimate covariances:** Covariances for this period are modeled separately for SR and NSR units (Section 3.3.4).
- **Pairs of years for which the average of correlations are estimated and modeled:** As described in Step 1 in Section 3.3.4, correlations between any 2 years in 2005–2015 are estimated and modeled. Then, the modeled (smoothed) correlation matrices are converted to covariance matrices based on the directly estimated variances.
- **Creating state-level covariance:** For any given state, the SR and NSR covariances are combined using the estimated sample size for the state (i.e., a weighted average based on the estimated sample size in the NR and NSR groups).
- **How this set is different from Set 1:** Because the covariances in 2004 and earlier (Set 1) are assumed independent from the covariances in 2005–2015 (Set 2), the covariance matrices are developed separately.

Set 3. SR and NSR covariances for years in 2014–2018

- **Method to estimate covariances:** Covariances for this period are modeled separately for SR and NSR units (Section 3.3.4).
- **Pairs of years for which the average of correlations are estimated and modeled:** Because the sample redesign occurred in 2016, the average of correlations is only estimated based on correlations between estimates in 2014 or 2015 and estimates in 2016, 2017, or 2018 (2014/2016, 2014/2017, 2014/ 2018, 2015/2016, 2015/2017, 2015/2018). Direct estimates of SR and NSR covariances are used for other pairs of years in this span of years.
- **Creating state-level covariance:** For any given state, the SR and NSR covariances are combined using the estimated sample size for the state (i.e., a weighted average based on the estimated sample size in the NR and NSR groups).
- **How this set is different from Set 2:** Because of the phase-in of the 2010 Census PSUs in 2016, this set is considered independent from Set 2. There is overlap between this

period (2014–2018) and the period in Set 2 (2005–2015), because if an estimate is requested for 2014–2018, it has units under both designs and is treated differently than an estimate based on years solely in 2015 and earlier.

Set 4. SR and NSR covariances separately for Group 1 and Group 2 for years 2016 and beyond

- **Method to estimate covariances:** The covariances are modeled separately for (a) SR, State Group 1; (b) SR, State Group 2; (c) NSR, State Group 1; and (d) NSR, State Group 2.
- **Pairs of years for which the average of correlations are estimated and modeled:** The averaging of correlations will not begin until 2016–2019, when the 2019 data become available. Before then, the directly estimated SR and NSR covariances for 2016–2018 are used within each group of states. Like the year 2005, 2016 will be dropped when averaging correlations because of the unusual way 2016 is estimated.
- **Creating state-level covariance:** The four covariances between SR and NSR are combined using the estimated population size (i.e., a weighted average based on population size within each group) instead of the estimated sample size as used for Sets 1–3. This step will be illustrated next.
- **How this set is different from Set 3:** In 2016, the NCVS design introduced state supplementation in the 22 largest states, which increased sample size in these states. This covariance estimation for this set accounts for the supplemental sample since 2016.

3.3.5.3 Creating State-Level Covariance for Set 4 (2016 and Beyond)

In Sets 1 and 2, the covariance matrices are combined using estimated sample sizes. Beginning in 2016, stratification was done within states, so actual sample sizes rather than expected sample sizes are used to combine covariance matrices. For Set 4, the SR proportion of the state population is estimated on a weighted basis,

$$f_i = \frac{\widehat{N}_{SR,i}}{\widehat{N}_{SR,i}+\widehat{N}_{NSR,i}}, \tag{3.5}$$

where $\widehat{N}_{SR,i}$ and $\widehat{N}_{NSR,i}$ are the weighted estimates of the SR and NSR populations in state i, averaged over years 2016 and beyond. The modeled state-level covariance is

$$C_i = \frac{n_{SR,g}}{n_{SR,i}} f_i^2 C_{SR,g} + \frac{n_{NSR,g}}{n_{NSR,i}} (1 - f_i)^2 C_{NSR,g}, \tag{3.6}$$

where $n_{SR,g}$ and $n_{NSR,g}$ are the sample sizes in the state group, g, to which state i belongs, and $C_{SR,g}$ and $C_{NSR,g}$ are covariances for SR and NSR areas in g.

3.3.6 Combining Covariances Across Sets

The SR and NSR covariances from Sets 1–3 are combined to produce state-level covariances running possibly as high as 2018. These state matrices are then integrated with the state-level matrices from Set 4 through the following steps:

1. Use the results from Set 4 for any variance or covariance between estimates for 2016 and beyond.
2. Adjust the covariance between an estimate in 2016–2018 and one in 2014–2015 to reflect the difference between the variance estimated for 2016–2018 based on Sets 1–3 and the variance estimated based on Set 4. In other words, the estimated covariance is multiplied by the square root of the ratio of the Set 4 variance to the Sets 1–3 variance. This approach produces a covariance estimate based on the correlation from Set 3 and the variance from Set 4.

3.4 Software Implementation

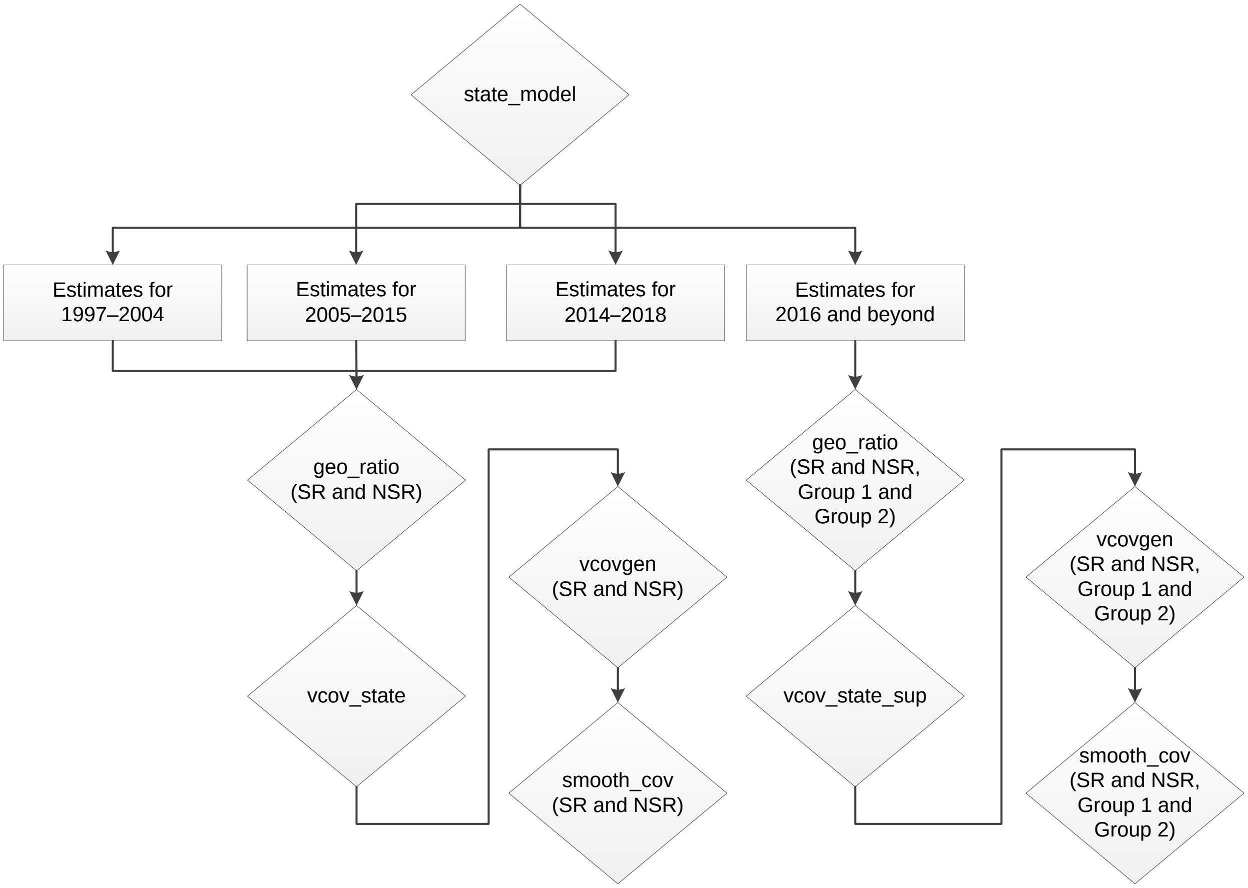
When using the NCVS SAE R programs to calculate the small area estimates (see discussion in Chapter 6), users will use the function *state_model.R*. This function in turn calls several other functions to estimate the covariance matrix given the model and years specified. The key functions that are used are as follows:

- *geo_ratios*: For Sets 1–3, it creates the linear substitutes for each year corresponding to the ratios being estimated (e.g., victimization rates or prevalence rates) separately for SR and NSR areas. For Set 4, it creates linear substitutes for SR and NSR areas in State Groups 1 and 2. The appropriate linear substitutes are passed to the function *vcov_state* for Sets 1–3 and to *vcov_state_sup* for Set 4 to calculate direct variance estimates for the ratios being estimated.
- *vcov_state*: It estimates SR and NSR covariance matrices when all years are in 2018 and earlier. This is a function specific to NCVS and estimates the covariance separately for different time spans to enforce assumed covariance structure with 0 covariance between detailed periods.
- *vcov_state_sup*: It estimates SR and NSR covariance matrices within Group 1 and Group 2 when years are 2016 and beyond. This is a function specific to NCVS.
- *vcovgen*: It estimates covariance matrix using linear substitutes for SR and NSR units separately. This is a general function in the sae2 package.
- *smooth_cov*: In this function, correlations are smoothed using regression as discussed previously. This is a function specific to the NCVS.
- Figure 3.1 displays the process by which the functions are implemented. These listed functions are called only within *state_model* in the SAE procedures. As discussed in the previous sections and shown in Figure 3.1, the years of estimates affect the estimation method. Because the *state_model* function can estimate the covariance matrix when conducting SAE modeling with the NCVS data from 1997 to 2018 and beyond, users do

not need to develop additional R programs for this calculation if the sample design in future years stays the same as the ones in 2016–2018. If the sample design changes in future years, statisticians should refer to Section 3.3 to modify the R functions listed above.

For estimating at the substate level (county or CBSA), the process is very similar, but the functions called are *substate_model*, *vcov_substate*, and *vcov_substate_sup*, which in turn call the general estimating functions of *vcovgen* and *smooth_cov*. R codes that create all the R functions can be found among the supplemental files.

Figure 3.1 Called-out Functions Under *state_model* Function to Calculate Covariance Matrix for the NCVS SAE Modeling



CHAPTER 4. USE OF AUXILIARY INFORMATION IN DEVELOPING THE SAE MODELS

4.1 Introduction

Achieving success when developing a small area estimation (SAE) model requires having useful auxiliary variables that are highly correlated with and, thus, predictive to the target outcomes to be estimated. The differences in the crime victimization rates between the National Crime Victimization Survey (NCVS) and the Uniform Crime Reporting (UCR) Program's Summary Reporting System (SRS) are well documented (Lynch & Addington, 2009; U.S. Department of Justice, 2004). By design, the UCR SRS data reflect only crimes reported to law enforcement agencies (LEAs), but the NCVS data can reveal the "dark figure" of unreported crime (Skogan, 1977).

Fay and Li (2011) investigated the relationship between the NCVS and UCR SRS rates at the county level, and Li et al. (2012) later examined this relationship at the state level. Both studies found that the long-term average of the NCVS county or state crime rates for each type of crime was best predicted by a single variable from the UCR SRS. Moreover, they found that in most instances the NCVS crime rate is best predicted by the corresponding rate from the UCR SRS (e.g., the NCVS robbery crime rate is best predicted by the UCR SRS robbery crime rate). Therefore, the UCR SRS variables are currently used in the SAE models as covariates to derive the NCVS small area estimates.

This section summarizes key results from Fay and Li (2011), Li et al. (2012), and Fay and Diallo (2015a) as they are related to the suitability of using the UCR SRS variables as predictors for the NCVS SAE models. Section 4.2 introduces components of the UCR SRS that closely parallel those of the NCVS and thus are considered as candidate predictors in the SAE models. It also summarizes some key results based on the evaluation of the UCR SRS data to facilitate determining proper model specifications for the NCVS SAE models, which eventually yielded the development of the multivariate dynamic model as described in Chapter 2. Section 4.3 provides a series of regression prediction analyses conducted at the county and state levels to identify the best predictor in the UCR SRS for each type of outcome used to determine the NCVS crime rate. Section 4.4 summarizes the final UCR SRS predictors selected for use in the NCVS SAE models.

Summary of Contents in Chapter 4

	Summary
What is the main point of this chapter?	This chapter summarizes the key results from the past work that led to the decision of using the UCR SRS variables as auxiliary data for the NCVS SAE models and selecting the predictors from UCR SRS data to model each type of NCVS crime rate. In the final section, suggestions are also given to readers who want to select predictors using NIBRS data and/or other auxiliary data sources for future NCVS SAE analyses.
Why is it important?	Achieving success when developing a SAE model requires having useful auxiliary variables that are highly correlated with and, thus, predictive to the target outcomes to be estimated. Identifying and evaluating potential auxiliary variables to be used as predictors is an essential step in the SAE analysis.
How is it operationalized?	The UCR SRS data are considered as auxiliary information for the NCVS SAE analysis because they are currently the most complete and predictive proxy of the underlying NCVS crime rates across different geographical locations and time. The UCR SRS data were evaluated to provide additional insight for determining the proper model specifications for the NCVS SAE models. Simple linear regressions of NCVS crime rates on the corresponding UCR SRS rates were conducted and the regression results proved that the UCR SRS rates have strong predictor power of the NCVS crime rates, at both state and county level. The regression results were also used to select final predictors of UCR SRS rates for each type of NCVS crime rate.

4.2 Auxiliary Information From the UCR SRS

4.2.1 Description of the UCR SRS and the NIBRS Data

The Federal Bureau of Investigation (FBI) through its UCR Program collects and reports statistics on the number of offenses known to law enforcement from participating LEAs across the entire nation. Two major systems under the UCR Program collect data on crimes in the United States: (1) the SRS and (2) the National Incident-Based Reporting System (NIBRS).

The traditional SRS was launched in the 1930s. In the SRS Return A form, LEAs tally the number of occurrences of eight crime types, known as Part I offenses or Type I offenses, which include the following:[10] (1) murder and nonnegligent homicide, (2) rape (legacy and

[10] Definitions of these crime/offense types are available at https://www.ucrdatatool.gov/offenses.cfm.

revised), (3) robbery, (4) aggravated assault, (5) burglary, (6) larceny-theft, (7) motor vehicle theft, and (8) arson. Then LEAs submit the aggregated counts of the collected data in monthly summary reports either directly to the FBI or indirectly through their state UCR programs. Because some crime incidents can concurrently have more than one crime type, the SRS simplifies the reporting by using a hierarchy rule that ranks the crime types in the order of severity and accepts the report of only the most severe crime type within a criminal incident. For example, if robbery, rape, and murder are committed within one crime incident, the agency will count this incident as one incident of murder given that murder is the most severe crime type among the three crime types. In SRS reports, ***violent crime*** covers four offenses: murder and nonnegligent homicide, rape, robbery, and aggravated assault; ***property crime*** includes the offenses of burglary, motor vehicle theft, larceny-theft, and arson.

To improve the overall quality of crime data collected by law enforcement, the UCR Program launched a new crime reporting system in 1990, known as NIBRS, that captures details on each single crime incident for 52 Group A offenses.[11] An increasing number of LEAs have transitioned from SRS to NIBRS in the past two decades. In 2018, approximately 44% of participating LEAs in the UCR Program submitted their data via NIBRS. However, the coverage of NIBRS data varies across different geographical locations. All of the LEAs in some states have become NIBRS reporters, while some states still have no LEAs reporting to NIBRS.
To maintain the completeness of the UCR SRS data in each year, the FBI applies the hierarchy rule used in the SRS to the NIBRS data and calculates the monthly crime counts for NIBRS reporters as their UCR SRS data. Therefore, the UCR SRS data have better coverage than NIBRS in all states and are considered as the auxiliary data source under the current NCVS SAE work.

It is worth noting that the UCR Program is planning to sunset the traditional SRS and transition to a NIBRS-only data collection by 2021. So, it is likely that analysts will need to consider using NIBRS data as the auxiliary data source for future NCVS SAE work after 2021. If that is the case, analysts have two options: (1) convert the NIBRS data into the UCR SRS data by applying the hierarchical rule, then continue to use the current SAE models for the NCVS; or

[11] See the current NIBRS user manual (FBI, 2020) for more details about the 52 Group A offenses: https://www.fbi.gov/services/cjis/ucr/data-documentation.**Error! Hyperlink reference not valid.**

(2) reselect useful auxiliary variables directly from the NIBRS data as covariates in the SAE models.

This report considers the UCR SRS data as auxiliary information for the NCVS SAE analysis because they are currently the most complete and predictive proxy of the underlying NCVS crime rates across different geographical locations and time. Section 5.2 provides details on how to access the UCR SRS data. Section 4.4 makes some recommendations on the model-building process when one wants to use the NIBRS or other auxiliary data for future NCVS SAE analyses.

4.2.2 Evaluation of the UCR SRS Data

Li et al. (2012) discussed three main sets of analyses they conducted to evaluate the components of the UCR Program's SRS data that closely parallel those of the NCVS. They also assessed the correlations of the SRS data with their counterparts in the NCVS data. Each of their sets of analyses has a different purpose and provides some insight for determining the proper model specifications for the NCVS SAE models.

4.2.2.1 Set 1: Examination of Geographic Patterns of Different Crime Types That Appear Distinctively Different

Key results: In the first set of analyses, the authors examined how state-level UCR SRS rates of the eight crime types during 2008–2010 can vary across the states. It was found that the geographic patterns of different crime types appear distinctively different. In other words, it is not always true for a state to have consistently higher or lower rates than the national average for all subtypes under violent crime or property crime. For example, the rate for robbery in Colorado was 50% below the national average for robbery, but the rate for rape was 50% above the national average for rape based on the 2008–2010 UCR SRS data.

Implication: The results suggested that "crime is a multi-dimensional phenomenon" (Li et al., 2012). Therefore, it is better to model each crime type separately and sum to total rather than model violent crime or property crime as a whole.

4.2.2.2 Set 2: Comparison of the Average Crime Rates Between 2008–2010 and 1998–2000

Key results: In the second set of analyses, the authors compared for each crime type the average crime rates (i.e., the average of the annual UCR SRS crime rates at the state level) for 2008–2010 with the average crime rates for 1998–2000 (i.e., exactly 10 years earlier). They found that the correlation between the 2008–2010 average crime rates and the 1998–2000 average crime rates was quite high even over a decade for each crime type. Furthermore, if State A had a higher average rate of a crime type than State B in 1998–2000, State A still had a higher average crime rate in 2008–2010 for this crime type.

Implication: This result implied stability in the relative ranking of states for each crime type over time.

4.2.2.3 Set 3: Comparison of the Crime Rates in the Four Largest States With the National Crime Rates in 1996–2010

Key results: In the third set of analyses, the authors compared the UCR crime rates in the four largest states (California, Texas, New York, and Florida) with the nationwide crime rates across time from 1996 to 2010. It was found that the trends in each state followed the national trends to a large extent for each crime type, which again suggested the stability of crime rates across time.

Implication: As found in the prior two sets of analyses (i.e., Sets 2 and 3), evidence of the stability of crime rates across time was utilized when developing the final NCVS SAE models on three different aspects:

- First, evidence of data stability backed the strategy that employs information from the NCVS sample values in neighboring years rather than just for the target year being estimated.
- Second, the evidence suggested incorporating time series into the models due to high correlations across time.
- Third, the evidence supported the model assumption that the geographic variation in the crime rates is relatively stable over time.

4.3 Regression Prediction of State- and County-Level NCVS Rates From the UCR SRS

As mentioned in Section 4.1, achieving success when developing an SAE model depends on how well the auxiliary variables can predict the target outcomes to be estimated. Fay and Li (2011) evaluated the predictive power of the UCR SRS rates on the corresponding NCVS rates when both were averaged over the same time period (e.g., from 1996 to 2005). Their analysis was mostly restricted to self-representing counties in the NCVS because they had relatively complete UCR SRS reporting with more stable results.

Fay and Li (2011) conducted simple linear regressions of NCVS crime rates on the corresponding UCR SRS rates. Table 4.1 shows the regression results for violent crime and its subtypes. For this analysis, Fay and Li (a) compared the UCR SRS aggregated and disaggregated rates of violent crime with aggregated NCVS violent crime rates and (b) compared the UCR SRS crime rates by disaggregated type of violent crime with NCVS crime rates disaggregated by type of violent crime.

Table 4.1 Regression Prediction of County-Level NCVS Violent Crime Rates From the UCR SRS Rates in Self-Representing Counties With the Highest Rates of Complete UCR SRS Reporting, for 1996–2005

Dependent Mean	NCVS Violent Crime	NCVS Violent Crime	NCVS Aggravated Assault	NCVS Rape/Sexual Assault	NCVS Robbery
Intercept	30.28[a]	30.28	6.05	1.25	4.29
UCR SRS Violent Crime	23.14 (1.36)[b]	18.89 (1.60)	3.20 (0.47)	0.59 (0.19)	0.98 (0.31)
UCR SRS Aggravated Assault	1.15 (0.19)	-1.14 (0.56)	0.00 (0.16)	-0.15 (0.07)	-0.19 (0.11)
UCR SRS Forcible Rape		31.04 (5.68)	8.49 (1.66)	3.36 (0.66)	0.43 (1.12)
UCR SRS Robbery		2.76 (0.67)	0.08 (0.20)	0.08 (0.08)	1.83 (0.13)

NCVS = National Crime Victimization Survey; SRS = Summary Reporting System; UCR = Uniform Crime Reporting.

[a] Beta coefficient for the intercept.

[b] Beta coefficient (standard error of beta coefficient).

Source: Fay and Li (2011), Table 4.

Key results: For the first comparison in Table 4.1, when using the UCR SRS violent crime rate to predict the NCVS violent crime rate, the beta coefficient is 23.14 with a standard error (SE) of 1.36. This result is statistically significant, meaning that reported violent crime in the UCR SRS has strong predictive power of total violent crime in the NCVS. However, when disaggregating the UCR SRS violent crime rate into its three components and individually using each component to predict the NCVS violent crime rate, the three components have very different prediction powers on the NCVS violent crime:

- Forcible rape from the UCR SRS has the highest coefficient, 31.04 (with an SE of 5.68), indicating a strong, statistically significant, correlation with the NCVS violent crime.
- Robbery from the UCR SRS has a coefficient of 2.76 (with an SE of 0.67), which, while smaller, is still statistically significant.
- Aggravated assault, surprisingly, has a coefficient of -1.14 (with an SE of 0.56), indicating a negative relationship with the NCVS violent crime.

For the second comparison, when using these three components to predict the corresponding three components of the NCVS violent crime rate, the UCR SRS robbery is the best single predictor of the NCVS robbery, and the UCR SRS forcible rape is the best single predictor of the NCVS rape/sexual assault. However, the UCR SRS aggravated assault is not a good predictor for the NCVS aggravated assault. Nevertheless, Fay and Li (2011) found that the UCR SRS forcible rape stood up as a strong predictor of the NCVS aggravated assault.

Li et al. (2012) refitted a series of similar regression models at the state level and found that the regression relationships between the UCR SRS components and the NCVS components are similar to what they found at the county level. They also evaluated the relationship of the three components under the property crime rate (i.e., burglary, motor vehicle theft, and larceny) between the UCR SRS and the NCVS. Each component in the UCR SRS was found to be the best predictor for its corresponding component in the NCVS.

Implications: The UCR SRS can be used to predict the NCVS for both aggregated crime types (i.e., violent and property crime) and disaggregated crime types (e.g., forcible rape, aggravated assault, burglary) at both the state and county levels.

4.4 Final Selected UCR SRS Predictor for Each Type of NCVS Crime Rate

Li et al. (2012) selected the best predictor from the UCR SRS for each of the NCVS crime rates using analyses described in Section 4.2. Table 4.2 summarizes the UCR SRS predictors used for each type of NCVS crime rate in the NCVS SAE analyses.

Key Result: For violent crime, the UCR SRS forcible rape was found to be the best predictor of rape and sexual assault, simple assault, and aggravated assault in the NCVS. The UCR SRS robbery was the best predictor for the NCVS robbery. For property crime, the UCR SRS motor vehicle theft, UCR SRS burglary, and UCR SRS larceny each can predict their analogous characteristics in the NCVS.

Table 4.2 UCR SRS Predictor for Each Type of NCVS Crime Rate

NCVS Rate	Best UCR SRS Predictor
Violent Crime	**Violent Crime**
Rape/Sexual Assault	Forcible Rape
Robbery	Robbery
Aggravated Assault	Forcible Rape
Simple Assault	Forcible Rape
Property Crime	**Property Crime**
Household Burglary	Burglary
Motor Vehicle	Motor Vehicle Theft
Theft	Larceny

NCVS = National Crime Victimization Survey; SRS = Summary Reporting System; UCR = Uniform Crime Reporting.
Source: Li et al. (2012), Table 3.

In addition, Fay and Diallo (2015a) discussed that the UCR SRS violent crime includes murder but murder is not included in the NCVS violent crime. Also, Fay and Li (2011) found that the UCR household tenure (i.e., whether the household owns or rents the house) did not show a strong relationship at the area level with renters reporting higher rates of violent crime in the NCVS. Therefore, these two variables (murder and household tenure) were not used as predictors in the final SAE models.

Implications: Although the NCVS rate can be predicted by the UCR SRS rate, the best predictor for the NCVS rate may not be the expected, corresponding, crime type in the UCR SRS. Each of the UCR SRS predictors listed in Table 4.2 is used for each of the types of the NCVS crime rates in the SAE analyses.

4.5 Suggestions for Selecting Predictors Using NIBRS Data and Other Auxiliary Data Sources for Future NCVS SAE Analyses

As discussed in Section 4.2.1, the traditional UCR SRS data collection may sunset after 2021, which means that NIBRS will be the only collection of administrative data of all crime types. If the UCR SRS becomes unavailable in the future, analysts will need to use other auxiliary data sources to find new predictors to be used in the NCVS SAE models. Analysts can adopt similar approaches as described in this section to select predictors among variables that are available in the NIBRS data or other auxiliary data sources (e.g., the decennial census data or American Community Survey data).

Here are some key takeaways from the previous research as described in Section 4.2 that are recommended to consider when analysts want to select new predictors from new auxiliary data:

- **Model each crime type separately rather than model violent crime or property crime as a whole.** As shown in Table 4.1, the predictors for different crime types can be different. Therefore, it is recommended to model each crime type separately using different predictors in the multivariate dynamic models.
- **Consider all possible candidate predictors in the initial model.** The “obvious” choice of predictor may or may not be the best choice. Some components of the UCR SRS crime rates are the best predictors of their corresponding components of the NCVS crime rates, and some are not. For example, the UCR SRS robbery is the best single predictor of the NCVS robbery, but the UCR SRS aggravated assault is not as good a predictor as the UCR SRS forcible rape variable for predicting the NCVS aggravated assault rate. Therefore, it is better to include all possible candidate predictors in the initial model, then conduct a variable selection procedure (e.g., backward or stepwise variable selection procedure) to select the best predictors.
- **Evaluate the regression relationships at both the county and state levels.** If analysts want to conduct both state- and county-level SAE analyses (or an SAE analysis at the Core-Based Statistical Area level), they need to confirm the regression relationships between the selected predictors and the NCVS components at each area level. If they find that the regression relationships vary across different area levels, they need to use different predictors in their SAE analyses for different area levels.

CHAPTER 5. DATA PROCESSING PROCEDURES TO PREPARE THE DATASETS FOR SMALL AREA ESTIMATION

5.1 Introduction

This chapter provides the detailed instructions for preparing datasets that will be used in the small area estimation (SAE) procedures for National Crime Victimization Survey (NCVS) described in Chapter 6. The three key pieces of data to be processed are (a) the Uniform Crime Reporting (UCR) Summary Reporting System (SRS) data (Section 5.2), (b) the NCVS data (Section 5.3), and (c) the census and the American Community Survey (ACS) data (Section 5.4). Auxiliary variables from the UCR SRS data are used as predictors in the SAE models, and the census and ACS data are used to calculate the state-level population estimates. The state-level population estimates, in turn, are used in the benchmarking procedures to adjust the state-level NCVS small area estimates to agree with their published national totals in NCVS.

Summary of Contents in Chapter 5

What is the main point of this chapter?	This chapter provides the detailed instruction on how to download, process, and generate datasets to be used in the NCVS SAE procedures. These datasets include (a) UCR SRS data, (b) NCVS data, and (c) census and ACS data.
Why is it important?	The datasets need be formatted in particular ways to be used in the corresponding R functions for the SAE analysis. In the SAE modeling process, variables in the UCR SRS data are used as predictors, and variables in the NCVS data are used as dependent (outcome) variables. The census and the ACS data are used to calculate the state-level population estimates for benchmarking the state-level small area estimates to meet with the national totals.
How is it operationalized?	• The UCR SRS crime rates at the state or substate level first are calculated based on the UCR SRS data and census population total data. The crime rates for each crime type are then exported into separate comma-separated values (CSV) files. • Victimization counts from the NCVS incident data files are tallied and merged onto the NCVS person- and household-level files to create person- and household-level NCVS data files for SAE modeling. • The state-level population estimates of the NCVS target population are calculated based on both the decennial census data and the ACS data. Two separate CSV files are generated, one (at the person level) containing the estimated number of persons in the target population and another (at the household level) with the 3-year ACS estimated number of non-vacant housing units.

5.2 Preparation of the UCR SRS Data

5.2.1 Sources to Download UCR SRS Data

UCR SRS data from the Federal Bureau of Investigation (FBI) are crime data reported by law enforcement agencies and serve as the key auxiliary data in the NCVS SAE models. The FBI is changing how crime data are published in two main ways: (a) the website to which the data are published and (b) the level of detail. First, historically, the UCR SRS data were published in the Inter-university Consortium for Political and Social Research (ICPSR), a repository of social science data maintained by the Institute for Social Research at the University of Michigan. This repository holds UCR SRS data from the 1970s through 2016. The FBI is now hosting and maintaining the data on its Crime Data Explorer (CDE)[12] website. Second, in the past, data were reported monthly by agencies, with counts for each crime type, to SRS. The FBI is transitioning to an incident-based reporting system, known as the National Incident-Based Reporting System (NIBRS), to which law enforcement agencies need to report details for each crime incident rather than aggregated monthly tallies of crimes. More details about the UCR SRS and the NIBRS data are given in Chapter 4.

The SAE models for the NCVS described in this report use the UCR SRS data; the SRS is currently the only auxiliary data source of crimes for all crime types outside of NCVS that is available for the entire nation. On the ICPSR, crime estimates from the UCR SRS data required for the SAE models are available through 2016 at the state level to produce state estimates and through 2016 (except for 2015) at the county level to produce county and Core-Based Statistical Area (CBSA) estimates. The files for the state-level crime estimates of the UCR SRS data now are also available on the CDE website in the same data format as on the ICPSR through 2018. These files are quite straightforward and include the year (from 1979 to the most current year); the state; the total population; and the number of crimes for violent, homicide, legacy rape, revised rape,[13] robbery, aggravated assault, property crime, burglary, larceny, and motor vehicle theft. The UCR SRS state-level estimates can also be extracted from the FBI Crime Data

[12] https://crime-data-explorer.fr.cloud.gov/

[13] In 2013, FBI changed the UCR definition of rape, as most agencies were excluding several sex offenses on the basis of their interpretation of the previous definition of “the carnal knowledge of a female forcibly and against her will.” The new definition is “Penetration, no matter how slight, of the vagina or anus with any body part or object, or oral penetration by a sex organ of another person, without the consent of the victim.” See https://ucr.fbi.gov/recent-program-updates/new-rape-definition-frequently-asked-questions for discussion.

Application Programming Interface (API)[14] and directly imported into R. Section 5.2.2 provides more details on how to extract these data from the API. However, the CDE website does not yet have the county-level estimates of UCR SRS. In addition, the FBI plans to sunset the UCR SRS after 2021 and transition to a NIBRS-only data collection. It is unclear whether the FBI will continue to publish UCR SRS data once all agencies are participating in the NIBRS. If the UCR SRS data are no longer produced, analysts will need to summarize NIBRS data themselves to get a data structure similar to that of the current UCR SRS data for use in the SAE models.

In summary, the main points from this subsection are as follows:

- The state-level crime estimates of the UCR SRS data are used as predictors in the NCVS SAE models to derive state-level estimates for NCVS. The county-level crime estimates of the UCR SRS data are used as predictors in the NCVS SAE models to derive estimates at the county level and Metropolitan Statistical Area level for NCVS.
- At the time of this writing, the state-level estimates through 2018 are available on CDE and can also be extracted from the FBI Crime Data API. Section 5.2.2 describes the three-step procedure to extract the UCR SRS state-level estimates from API, then calculate the state-level crime rates based on these estimates, and finally generate input datasets that contain the calculated crime rates to be used in the SAE modeling procedure for state-level estimation.
- At the time of this writing, the county-level crime estimates of the UCR SRS data are available on ICPSR only through 2016 (except for 2015) and are not available on CDE. Section 5.2.3 describes the three-step procedure to generate the datasets for county-level estimation. This procedure is similar to the three-step procedure for the state-level crime estimates: processing the data files of the UCR SRS county-level estimates, then calculating the county-level crime rates based on these estimates, and finally generating input datasets to be used in the SAE modeling procedure for county-level and CBSA-level estimation.
- The FBI plans to sunset the UCR SRS after 2021 and transition to a NIBRS-only data collection. If the SRS data are no longer produced, analysts will need to summarize NIBRS data themselves to get a data structure similar to that of the current UCR SRS data for use in the SAE models.

[14] See https://crime-data-explorer.fr.cloud.gov/api for more information. First-time users of this API need to request an API key from https://api.data.gov/signup/.

5.2.2 Procedure to Prepare Datasets of the UCR SRS State-Level Estimates for the NCVS SAE Modeling Process

This subsection illustrates the three-step procedure that should be done to obtain the UCR SRS state-level estimates and convert them into datasets that will be used as input in the NCVS SAE modeling process. The supplemental file *AssembleUCR_State_Reproducible.R* provides the sample R code that extracts state census population totals from the Census Bureau's API[15] and the state-level UCR SRS estimates from the FBI Crime Data API. Although these data can also be downloaded from the Census Bureau's website and the CDE website, using the APIs to get the data is more efficient and is easier to reproduce. The state census population totals will be used to calculate the covariance matrix in the SAE procedure as illustrated in Chapter 3. Table 5.1 presents an overview of this procedure. Details of the three steps are discussed in full below the table.

Table 5.1 An Overview of the Procedure That Prepares Datasets of the UCR SRS State-Level Estimates for the NCVS SAE Modeling Process

	Description
Purpose	• This procedure produces datasets of the UCR SRS state-level estimates that will be used as input in the NCVS SAE modeling process to generate small area estimates at the state level.
R Program File	• *AssembleUCR_State_Reproducible.R*
Input Data Files	• Extract the state census population totals from the Census Bureau's API • Extract the state-level UCR SRS estimates from the FBI Crime Data API
Output Files	• Separate CSV files for each criminal offense are created to contain state-by-year crime rates (e.g., "*Robbery.est.csv*" for robbery) • A CSV file is created for state-by-year population sizes from the UCR SRS data ("*Population.est.csv*")
Tips to Check the Final Output Files	• To ensure that the output files are generated properly, check that each file contains a record for each state with variables State (state name); pop2010 (2010 population); fips (state FIPS code); stabbr (state abbreviation); and yr96, yr97, yr98, …, yr17, yr18 (the annual crime rates per 100,000 persons)
Future-Year Changes	• Make sure all the variable names remain the same in the future-year data • Update the values in "startyear" and "endyear" in the R program to define the year period of interest

CSV = comma-separated values; FIPS = Federal Information Processing Standards.

[15] The U.S. Census website offers information on available APIs at https://www.census.gov/data/developers/data-sets.html. The first-time user of Census API needs to request a key from the website. See https://www.census.gov/data/developers/guidance/api-user-guide.html for more details. This API can be accessed directly through R using the *tidycensus* package.

5.2.2.1 Step 1—Import the UCR SRS State-Level Estimates and the State-Level Population Estimates Into R

This step involves importing the UCR SRS state-level estimates and the state-level population estimates into the R program. The key elements in this step are as follows:

- **Import UCR SRS state-level estimates:** The state-level estimates of UCR SRS data can be downloaded from the CDE website. This file is named "estimated_crimes.csv" under the "Summary (SRS) Data with Estimates" section on CDE's "Documents & Downloads" webpage. To date, this file contains the national- and state-level estimates of UCR SRS data as well as the population size data for each year from 1979 through 2018. FBI will continuously update this data file to include estimates from the most recent year. This file has a record for each combination of state and year, and it contains variables indicating population size and the estimated state-level counts of homicide, rape, legacy rape, robbery, aggravated assault, property crime, burglary, larceny, and motor vehicle theft at each year time point. Alternatively, the UCR SRS state-level estimates and population sizes can be extracted from the FBI's Crime Data API, which is the method demonstrated in the supplemental files. The UCR SRS data file downloaded from the CDE website contains not only the state-level estimates, but also the national-level estimates. The row for the national-level estimates should be removed for the use of state-level SAE analysis.
- **Download 2010 state census population total data file:** The 2010 (or the most recent decennial census year) state census population totals can be downloaded from the U.S Census Bureau website or extracted from its API, as demonstrated in the supplemental files.

5.2.2.2 Step 2—Calculate the State-Level Crime Rates Based on the UCR SRS State-Level Estimates

This step involves calculating the state-level crime rates per 100,000 persons. The key elements in this step are as follows:

- **Calculate crime rates:** The state crime rates per 100,000 persons are calculated by multiplying the count of each criminal offense by 100,000, then dividing by the state population size. All the variables, including crime count and population size, are from the UCR SRS data.
- **Understand the difference in how rape is captured (legacy rape or revised rape):** The two rape variables in the UCR SRS data are "rape_legacy" and "rape_revised." The variable rape_legacy represents only female victims and excludes sodomy and sexual assault with an object. This variable is present up until 2016. The variable rape_revised adds to rape_legacy by including male victims, sodomy, and sexual assault with an object. This variable was added to the NIBRS in 2013.

- **Modify the legacy rape variables for 2017 onward:** Using the 4-year overlap period in which the NIBRS included both rape_legacy and rape_revised, an adjustment factor between the two measures was calculated by taking the average slope of the 2013, 2014, and 2015 regressions[16] of rates of legacy rates upon revised rapes on the basis of the UCR SRS data. The resulting adjustment factor is 0.723. Using this adjustment factor, starting in 2017, the legacy rape estimates are estimated by multiplying the revised rape estimates by the 0.723 adjustment factor.

5.2.2.3 *Step 3—Export State-Level Crime Rates and Census Population Estimates Into Comma-Separated Values Files*

This step involves merging the calculated crime rates from Step 2 with the 2010 state census population total data from Step 1, and then generating the final output files to be used in the NCVS SAE modeling process to generate state-level small area estimates for NCVS. The key elements in this step are as follows:

- **Export crime rates of each criminal offense into separate comma-separated values (CSV) files:** The 2010 state census population total file from Step 1 is merged with crime rates of each criminal offense at the state-by-year level. Then, a dataset is exported as a CSV file for each criminal offense, including burglary, larceny, motor vehicle theft, legacy rape, and robbery. For example, the CSV file for robbery is named "*Robbery.est.csv*." These CSV files have a record for each state. Each file contains the variables State (state name); pop2010 (2010 population); fips (state Federal Information Processing Standards [FIPS] code); stabbr (state abbreviation); and yr96, yr97, yr98, …, yr17, yr18 (the annual crime rates per 100,000 persons for the corresponding crime).
- **Export the 2010 Census population estimates into a CSV file:** The 2010 state census population total file from Step 1 is merged with population sizes in the UCR SRS state estimate data at the state-by-year level. This dataset is then exported as a CSV file named "*Population.est.csv*." The file contains one record for each state. Each record contains the variables State (state name); pop2010 (2010 population); fips (state FIPS code); stabbr (state abbreviation); division (census division code); and yr96, yr97, …, yr17, yr18 (the annual population of the state from the FBI UCR file).

[16] Data for 2016 were not available at the time this decision was made, but they have a similar slope.

5.2.3 Procedure to Prepare Datasets of the UCR SRS County-Level Estimates for the NCVS SAE Modeling Process

This subsection illustrates the three-step procedure to obtain the data files with the UCR SRS county-level[17] estimates from the ICPSR website and convert them into datasets that will be used as input in the NCVS SAE modeling process to generate small area estimates at the county or CBSA level. This procedure is similar to the way in which the state estimates are obtained as mentioned in the previous subsection—with some deviations because the county-level estimates are not available at the CDE website or FBI Crime Data API. Table 5.2 presents an overview of this procedure. Details of the three steps are discussed in in full after the table.

Table 5.2 An Overview of the Procedure That Prepares Datasets of the UCR SRS County-Level Estimates for the NCVS SAE Modeling Process

	Description
Purpose	• Produces datasets of the UCR SRS county-level and CBSA-level estimates that will be used as input in the NCVS SAE modeling process to generate small area estimates at the county or CBSA level
R Program Files	• County-level estimates: *kaplan2x.R* • CBSA-level estimates: *CBSA.mich.R* • Additional function to read in the datasets: *ucr_counties_rev.R* • Create crosswalks between CBSA and counties: *CBSA.def.rev.R*
Input Data Files	• Annual data files of UCR SRS county-level estimates from ICPSR (named “Uniform Crime Reporting Program Data: County-Level Detailed Arrest and Offense Data, United States”) • *statestub.csv*: 2010 state census population totals downloaded from the U.S Census Bureau website • *co-est2018-alldata.csv**: 2018 (most recent year) ACS county-level population estimates data from the Census Bureau
Output Files	• Separate comma-separated values (CSV) files for each criminal offense are created to contain county-by-year crime rates (e.g., “Robbery.est.csv” for robbery) • CSV files are created for CBSA-by-year crime rates of each criminal offense (e.g., “*Robbery. cbsa.csv*” for robbery) • A CSV file is created for county-level population estimates for counties to be modeled (“*county.data.csv*”)
Tips to Check the Final Output Files	• To ensure that the output files are generated properly, check that each file has a record for each county or CBSA and contains variables as listed in Step 3

[17] The term “county” includes county-equivalent areas as defined by Census bureau FIPS codes, including traditional counties; boroughs in Alaska; parishes in Louisiana; independent cities in Maryland, Missouri, Nevada, and Virginia; and Washington, DC.

	Description
Changes to Be Made in Future Years	• It is unclear whether the ICPSR will continue providing UCR SRS county-level estimates for years after 2016. If yes, the R codes can be modified to take in new data from future years. If not, users need to download the Return A Master File from the CDE website and calculate county-level estimates themselves.

Note. *This file can be downloaded from the Census Bureau’s website: https://www2.census.gov/programs-surveys/popest/datasets/2010-2018/counties/totals/.

5.2.3.1 Step 1—Download and Import the UCR SRS County-Level Estimates From ICPSR, and County-Level Population Estimates From the U.S. Census Bureau Website, Into SAS and R

This step involves downloading and importing two sets of input data files into the R program: the annual UCR SRS county-level estimate data files and the census population total data. The key elements in this step are as follows:

- **Download UCR SRS county-level estimate data files from ICPSR:** The annual UCR SRS county-level estimate data files can be downloaded on the ICPSR website for the years from 1997 to 2016, except for 2015, resulting in 18 separate files. These files are named “Uniform Crime Reporting Program Data: County-Level Detailed Arrest and Offense Data, United States” on ICPSR. Data are not yet available for 2015. On ICPSR, these files are all in ASCII fixed-width file format. Programs originally in SAS and more recently in R are also provided to convert these files into CSV format. These programs have been modified and saved as supplemental files, including *icpsr-setup-1994-1999.sas* (for 1994–1999 data), *icpsr-setup-2008.sas* (for 2000–2008 data), *icpsr-setup-2009.sas* (for 2009–2010 data), icpsr-csv.sas (to convert the outputs from the previous three programs to CSV files) and *UCR_counties_update.R* (for 2011–2014 and 2016). The output data sets have names such as ucr97.csv, etc. A file urc16.csv is included as a supplement file to demonstrate the format of the final output data file. Note that a file ucr15.csv was created by copying the data in ucr16.csv. The 2015 file should be updated by running *UCR_counties_update.R* once the 2015 county-level estimate data become available on ICPSR.
- **Download the annual county-level population data and the 2010 state census population data:** The ACS data file with county-level population, migration, birth, and death data for the most recent year (i.e., 2018), named *co-est2018-alldata.csv*, can be downloaded from the U.S. Census Bureau website. The 2010 state census population totals (*statestub.csv*) should also be downloaded. The 2015 CBSA definitions used by the 2017 ACS are used and saved in *list1_Jul_2015.csv* to calculate the CBSA-level crime rates. These definitions can also be downloaded from the Census Bureau’s website.[18]

[18] https://www.census.gov/geographies/reference-files/time-series/demo/metro-micro/delineation-files.html

5.2.3.2 Step 2—Calculate the County-Level and CBSA-Level Crime Rates Based on the UCR SRS Estimates

This step involves calculating the county-level and CBSA-level crime rates per 1,000 persons (not 100,000 as for state-level crime rates in the previous section). The key elements in this step are as follows:

- **Identify the counties for which small area estimates can be produced:** The SAE procedure cannot include all counties listed in the UCR SRS county-level data. To warrant the accuracy of the final small area estimates, only counties that have adequate sample sizes in NCVS and have relatively complete UCR SRS reporting are selected to generate their county-level small area estimates. There are 65 counties meeting these criteria. The FIPS codes of these 65 counties are listed in the *modeled.counties.2.csv* file. Note that the FIPS code 36061 for New York county was used to represent all five boroughs of New York City.
- Calculate annual crime rates at the county level: The annual UCR SRS county-level estimate data files are used to calculate the annual crime rates for each criminal offense type at the county level for each of the 65 counties. To get the crime rate for each criminal offense at the county level, the county-level UCR SRS estimate of the count of each criminal offense is multiplied by 1,000 and then divided by the county-level population size, which is also provided in the annual UCR SRS county-level estimate data.
- Identify CBSAs for which small area estimates can be produced: The goal of the NCVS SAE project is to include all of the 53 CBSAs that have at least 1 million population. The R program *CBSA.def.rev.R* creates a hard-coded list of 53 CBSAs using the crosswalk data file (named *list1_Jul_2015.csv* in the supplemental file) of the 2015 CBSA definitions used for the 2017 ACS. The 2015 CBSA definitions link each CBSA code with the FIPS codes of all its counties. The output data file from this program is also included as a supplemental file named *CBSA_def_2017.csv*. This output file includes a variable named “main.state,” indicating the state with the largest proportion of the population for each CBSA. This variable is used in modeling covariance matrices, as discussed in Chapter 3. This program also updates the CBSA code for Los Angeles from the old code 31100 to the new code 31080 that went into effect 2013.
- Calculate annual crime rates at the CBSA level: In a way similar to that in which the county-level crime rates are calculated, the annual UCR SRS county-level estimate data files are used to calculate the annual crime rates for each criminal offense type at the CBSA level. With the CBSA_def_2017.csv file that provides crosswalks between the CBSA code and FIPS codes of all its counties, the count of each criminal offense for each CBSA is the sum of its county-level UCR SRS counts, and the population size of each CBSA is the sum of its county-level population sizes in the UCR SRS data. The CBSA-level crime rate is calculated by multiplying the CBSA-level count by 1,000 and then dividing by the CBSA-level population size.

- Use the 2016 UCR SRS data as the 2015 data: Because the UCR SRS county-level estimate data file is not available for 2015 on ICPSR, the 2016 UCR SRS county-level estimate data file is used to calculate the crime rates for 2015. Therefore, the crime rates of 2016 and 2015 are the same for each selected county and CBSA.

5.2.3.3 Step 3—Export the County-Level and CBSA-Level Crime Rates and Census Population Estimates Into CSV Files

This step involves exporting the calculated crime rates from Step 2 and the census population estimates into CSV files. The key elements in this step are as follows:

- **Export the county-level crime rates of each criminal offense into separate CSV files:** The R program, *kaplan2x.R*, calculates the crime rates at the county level (in Step 2) and exports the crime rates as CSV files for each individual criminal offense, including aggravated assault, rape, robbery, burglary, larceny, and motor vehicle theft. For example, the CSV file for robbery is named "Robbery.co.csv." These CSV files have a record for each county. Each file contains the following variables: FIPS_ST (the two-digit state FIPS code); FIPS_CTY (the three-digit county FIPS code); FIPS (the five-digit state/county FIPS code); CTYNAME (county name); and yr97, yr98, …, yr15, yr16 (the annual crime rates per 1,000 persons). Note that the *kaplan2x.R* file calls an function in another R file, *ucr_counties_rev.R*, to read in the datasets properly.
- **Export the CBSA-level crime rates of each criminal offense into separate CSV files:** The R program *CBSA.mich.R* calculates the CBSA-level crime rates (in Step 2) and exports the crime rates as CSV files for each individual criminal offense, including aggravated assault, rape, robbery, burglary, larceny, and motor vehicle theft. For example, the CSV file for robbery is named "*Robbery.cbsa.csv*." These CSV files have a record for each CBSA. Each file contains the following variables: CBSA.code (CBSA code); yr97, yr98, …, yr15, yr16 (the annual crime rates per 1,000 persons); CBSA.Title; and main.fips (the two-digit state FIPS code of the state with the largest proportion of the CBSA's population for use in modeling the covariance matrix). Note that the *CBSA.mich.R* file also calls an function in the R file *ucr_counties_rev.R* to read in the datasets properly.
- **Compile the county-level population data using *county.data.R*:** The program *county.data.R* reads and combines the file *modeled.counties.2.csv*, the file *co-est2018-alldata.csv* with recent population estimates for the counties, and the file *statestub.csv* with state-level characteristics. This R program outputs the file *county.data.csv* for the set of 65 counties to be modeled. To change the set of modeled counties, *modeled.counties.2.csv* should be updated before *county.data.R* is run.

5.3 Preparation of the NCVS Data

This section describes (a) the processing of the NCVS data beginning with a description of the data and the timeliness and (b) the steps to process the data for the SAE modeling. The process for reading and preparing NCVS data for SAE is a multistep process that begins with NCVS data (including household-, person-, and incident-level files) provided by the Census Bureau and ending with a summary data file. This work must be done either at the Census Bureau or at a Census Research Data Center (RDC)[19] because of confidentiality requirements. The NCVS data are typically available by the April or May following the reporting year. For example, the 2018 NCVS data were available in May 2019. The steps outlined below are current as of NCVS data from 2004 through 2018. Every 10 years (years ending in 6), the NCVS has a new sample. Sometimes the new sample coincides with a major design change. For those pivotal years, this process may need to be revisited and modified. For example, in 2016, a special bridge file was constructed to bridge between the 2015 and 2016 data. BJS also plans a redesign of the NCVS instrument in the 2020 decade that may affect variables and variable names. Table 5.3 presents an overview of the procedures that prepare datasets of the NCVS summary data for the NCVS SAE modeling process. Details of the three steps involved in these procedures are described after the table.

Table 5.3 An Overview of the Procedure That Prepares Datasets of the NCVS Summary Data for the NCVS SAE Modeling Process

	Description
Purpose	• Puts the NCVS data into a format for SAE modeling. Victimization counts from incident files are tallied and merged onto the person- and household-level file for relevant crime types. This is done in 3 steps: (a) read in quarterly NCVS SAS data files, select relevant columns, and save as CSV file; (b) read in CSV files and save as RData objects; and (c) create victimization counts and save in person- and household-level files.
Programs	• *extract2009.sas, extract2010.sas* (Step 1) • *csv_to_Rdata_SAS.R* (Step 2) • *toc_prev_sas4.R* with *sasncvs_recode.R* and *relative.R* as source files (Step 3)

[19] https://www.census.gov/fsrdc

	Description
Input Data Files	• *ncvsYYYYq1hh.sas7bdat*, *ncvsYYYYq2hh.sas7bdat*, *ncvsYYYYq3hh.sas7bdat*, *ncvsYYYYq4hh.sas7bdat* • *ncvsYYYYq1per.sas7bdat*, *ncvsYYYYq2per.sas7bdat*, *ncvsYYYYq3per.sas7bdat*, *ncvsYYYYq4per.sas7bdat* • *ncvsYYYYq1inc.sas7bdat*, *ncvsYYYYq2inc.sas7bdat*, *ncvsYYYYq3inc.sas7bdat*, *ncvsYYYYq4inc.sas7bdat* • where YYYY is the 4-digit year (e.g., 2009); all these files are provided by the Census Bureau
Output Data Files Step 1	• sasncvsYYh.csv, sasncvsYYp.csv, sasncvsYYi.csv • where YY is the 2-digit year (e.g., 09)
Output Data Files Step 2	• *sasncvsYY.Rdata* (a R list with 3 data frames [i.e., datasets]—1 each for the household-, person-, and incident-level data)
Output Data Files Step 3	• *sastoc_per0810.Rdata*, *sastoc_prop0810.Rdata* • *sastoc_per1113.Rdata*, *sastoc_prop1113.Rdata* • *sastoc_per1416.Rdata*, *sastoc_prop1416.Rdata* • where the "per" files are person-level crime and the "prop" files are household-level crime and the ending digits refer to the years included in the file; for example, 0810 has data for 2008–2010 • Each file is an R list with a data frame for each year included; e.g., *sastoc_per0810.Rdata* contains 3 data frames—1 for each of years 2008, 2009, and 2010 • Each person-level data frame includes records for each of the sampled persons within each responding household, and each household-level data frame includes a separate record for each sampled household for each year
Tips to Check the Final Output Files	• For each year, calculate the estimated number of victimizations for each crime type by summing the type of crime variable. Compare this estimate to NCVS public criminal victimization reports, which should match (note that the NCVS reports round to the nearest 10).
Changes to Be Made in Future Years	• Incorporate more years of NCVS data • Check that variable names have not been changed—if they have changed, rename variables in Step 1 to be consistent with historic data

Note. CSV = comma-separated values.

5.3.1.1 Step 1—Read in NCVS Data File From SAS Data to CSV

The NCVS data files in the RDC are available as SAS datasets. For each year, the NCVS data are released in 12 data files—a household-level file, a person-level file, and an incident-level file for each quarter. A SAS macro has been developed to read in and put together a year's worth of data for each file type. For a given year, all four quarters of data are read in for the household level data and several variables are kept. A similar process is done for the person- and incident-level files.

The data are then saved as three CSV files each year, one for each record type. To date, this process has been done for each year of data from 1998 to 2018 in separate programs. The data are saved as CSV files for later use in R. An example program of this process in the supplemental files is *extract2009.sas*, which illustrates this process for 2007, 2008, and 2009.

5.3.1.2 Step 2—Convert CSV Data to RData

For each year, the CSV files are read into R, and then the three files (household, person, and incident data files) for each year are combined into an R list[20] that contains three data.frame objects in R. Then the R list is saved as an RData file, one for each year. The program to complete this process for years 1998–2018 is included in the supplemental file named *csv_to_Rdata_SAS.R.*

5.3.1.3 Step 3—Create Type of Crime Data for Each Year

Preparing the NCVS data files for estimation is a complex task. The NCVS Variance User's Guide (Chapter 2)[21] describes in detail how to prepare the files for estimation, and this process is generally followed in this program. This step creates the final type of crime (TOC) data at both the person and household level for each year in RData format to be used as input data in the SAE procedures, as described in Chapter 6. At a high level, the process does the following.

Summarize, rename, and merge the incident-level data with the person- and household-level files for each year: The incident-level files have a record for each incident, which means some household or person respondents with no reported incidents will not be on the file and some person or household records are included multiple times, once for each reported incident. The incident files thus need to be summarized at the person and household levels to be merged with the person- and household-level files for estimating victimization rates. The

[20] R list is an object in R which can contain elements of different types, such as numbers, strings, vectors, matrices and embedded lists.

[21] See https://www.bjs.gov/content/pub/pdf/NCVS_Variance_User_Guide%2011.06.14.pdf.

victimizations are summed using the series adjustment of the victimization weight.[22] In addition, many of the variables are renamed to have the same names as in the public data file.

A function (*toc.form*) is written to perform the following steps for a given year of data:

1. Rename variables to be consistent with the public-use file.
2. Summarize person-level incidents and merge onto person-level data.
3. Summarize household-level incidents and merge onto household-level data.
4. Save the merged data.

The detailed steps are as follows:

- **Subset the incident data:** The incidents files are subset to exclude incidents occurring outside the United States (V4022 != 1) and incidents whose victimization weight (WGTVICCY) is 0.
- **Adjust the victimization weights to account for multiple victimizations corresponding to one person or household:** The NCVS questionnaire collects data on series crime. Series crimes are incidents of a similar nature that happen more than six times for which the respondent cannot recall dates and other characteristics of the incidents. Each series crime is represented as a single record, and therefore the weight is adjusted to account for the multiple victimizations as follows:

    1. If not a series crime, then use WGTVICCY.
    2. Else If a series crime and occurs 6–10 times, use WGTVICCY*(number of times).
    3. Else If a series crime and occurs more than 10 times, use WGTVICCY*10.
    4. Else If a series crime and occurs an unknown number of times, use WGTVICCY*6.

- **Calculate weighted counts of victimizations for violent crimes and personal nonviolent crimes at the person level:** After the summarized incident-level data are merged with the person-level data, the weighted count of victimizations for each violent crime and nonviolent personal crime is calculated at the person level in each year quarter using the adjusted victimization weights (referred to as “the adjusted WGTVICCY”). Table 5.4 shows the definition of each personal crime outcome that is defined based on the TOC Code variable in NCVS (V4529)[23] and the victim-offender relationship. It is important to note that the NCVS conducts

[22] Series crimes are incidents of a similar nature that happen six or more times and for which the respondent cannot recall dates and other characteristics of the incidents. Each series crime is represented as a single record and the weighting counts series victimizations as the actual number of victimizations reported by the victim, up to a maximum of 10.

[23] This variable indicates the types of crimes experienced. Codes 1 through 20 represent violent crimes and 31 to 59 are property crimes. Codes 21, 22, and 23 include purse snatching and pick-pocketing, which are not considered to be either violent crimes (because they do not involve the use of threat or force) or property crimes (which are defined here as committed against a household).

interviews every 6 months, so a person or household can have two interviews in any given year.[24] In the NCVS data files, each interview record can be uniquely identified by the ID (for the person or household) along with the year quarter, and thus the weighted counts of victimizations are calculated for each person in each year quarter.

Table 5.4 Personal Crime Definitions

Derived Variable Name	Concept	Levels of V4529 (Type of Crime)	Other Conditions
rape	Rape/sexual assault	1, 2, 3, 4, 15, 16, 18, 19	
robbery	Robbery	5, 6, 7, 8, 9, 10	
sim.asslt	Simple assault	14, 17, 20	
aggr.asslt	Aggravated assault	11, 12, 13	
oth.personal	Nonviolent personal crimes (personal theft)	21, 22, 23	
ipv	Intimate partner violence	1 through 20	A spouse, ex-spouse, boyfriend, girlfriend, or ex-boyfriend or ex-girlfriend was the offender*
stv	Stranger violence	1 through 20	Offender was a stranger or would only be recognized by sight
otv	Other offender violence	1 through 20	Not ipv or stv

Note. *BJS' definition of the relationship of the offender to the respondent is complex. The program *relative.R* provides a function to define the relationship variable.

- **Calculate weighted counts of victimizations for property crimes using the household-level files:** After the summarized incident-level data are merged with the household-level data, the weighted count of victimizations for each property crime is calculated at the household level in each year quarter using the adjusted victimization weights. Table 5.5 shows the definition of each household crime outcome that is also defined by the TOC Code variable in NCVS (V4529).

Table 5.5 Household Crime Definitions

Derived Variable Name	Concept	Levels of V4529 (Type of Crime)
burglary	Burglary/trespassing	31, 32, 33
auto.theft	Motor vehicle theft	40, 41
larceny	Other theft	54, 55, 56, 57, 58, 59
total.prop	All property crime	31, 32, 33, 40, 41, 54, 55, 56, 57, 58, 59

[24] In 2016, the dataset was created using a bridge file that also included records from 2015, so some households and persons have three records on the 2016 file.

- **Calculate an indicator of any victimizations in each year for each person or household for estimating prevalence rates:** The weighted sums using the adjusted WGTVICCY are an annual estimate of the number of victimizations in a year and thus are used to estimate victimization rates. The prevalence rate, which is the percentage of persons who are victims of crime, is also of interest. An indicator of any victimizations in the year for each person is calculated to be able to estimate prevalence rates for each TOC.
- **Generate two R data frames—one for persons and one for households—to be used as input data in the estimation procedures:** The personal-level data frame contains a record for each person in each year quarter, along with columns of the weighted victimization sums and indicators for whether the person was victimized in the year for personal crime outcomes (as listed in Table 5.4). The *add.reltot* function in *relative.R* defines the relationship between the victim and the offender, which is used in some outcomes. The household-level data frame contains a record for each household in each year quarter, along with columns of the weighted victimization sums and indicators for whether the household was victimized in the year for household crime outcomes (as listed in Table 5.5). These two data frames are bundled together into a list in the function *toc.form*.
- **Run the estimation function for each year and save the results to disk:** The *toc.form* function has a year as an input. The function is run multiple times, once for each of several years, and multiple years' output are bundled together and saved as a RData file. The person-level outputs are saved together in one file and the household-level outputs are saved together in a separate file. Supplemental files *toc_prev_sas4.R*, *relative.R*, and *sasncvs_recode.R* are provided to accomplish this step.

5.4 Calculating State Controls for the Benchmarking Procedure Using the Census and ACS Data

The calculation of state-level small area estimates of crime rates involves the state-level population estimates. It is desirable that the state-level population estimates in the SAE process match corresponding population estimates based on both the decennial census data and the ACS data, as described below. To achieve this goal, benchmarking is performed (see Chapter 7 for more details on the benchmarking procedure). County and CBSA estimates are not benchmarked.

As stated previously, the NCVS is a survey of persons in the United States aged 12 and older who do not live in institutionalized group quarters. To calculate the population estimates for that specific population, the population estimates from both the decennial census and the ACS are used in conjunction to compute the population estimates for the NCVS target population at the state level. These estimates are used as the "gold standards" of the state population totals, referred to as *State Controls*, in the benchmarking procedure. The State Controls are calculated for both number of households and number of persons. Table 5.6 presents an overview of this procedure. Details for the three steps are discussed in in full after the table.

Table 5.6 An Overview of the Procedure That Calculates State Controls for the Benchmarking Procedure Using the Census and ACS Data

	Description
Purpose	• Calculates the State Controls at both household and person levels to be used in the benchmarking procedure
R Program File	• For person level: *state_pop_controls_reproducible.R* • For household level: *state_hh_controls_ reproducible.R*
Input Data Files	• Directly extract variables from Census Bureau's API
Output Files	• For person level: a comma-separated values (CSV) file with a record for each state and a column for each year from 2007 to 2018 containing the estimated number of persons aged 12 and older not living in institutionalized group quarters • For household level: a CSV file with a record for each state and a column for each year from 2007 to 2018 containing the 3-year ACS estimate of the number of non-vacant housing units
Tips to Check the Final Output Files	• To ensure that the output files are generated properly, check that each file has a record for each state and contains estimates for each year from 2007 to 2018
Changes to Be Made in Future Years	• Make sure the variable names remain the same in the future-year data • Update the values in "startyear" and "endyear" in the R program to define the year period of interest

5.4.1.1 Step 1—Import State Controls Data Into R

The first step is to download the population estimates of the decennial census and the ACS from the U.S. Census Bureau's website (data.census.gov). The 2010 Census data are used as the decennial census data in the current SAE work. The supplemental file *state_pop_controls_reproducible.R* provides the sample R code that extracts all the variables needed from both the decennial census and the ACS directly from the Census Bureau's API. The key elements in this step are as follows:

- **Download the 2010 Census summary data files:** Download the census summary files of population living in group quarters (PCO1) and institutionalized persons (PCO2) by age and state from the U.S. Census's website or request them directly from the API. The variable names in the API census data are different from the ones in the directly downloaded data from the website. A codebook for the API census data is at https://api.census.gov/data/2010/dec/sf1/variables.html. In the

API, the variables are P029026 and P029027 for the total in group quarters and the total in institutionalized group quarters, respectively.

- **Download the ACS data files:** The ACS estimates of each state's population (**B01003**); population under 18, which includes population by specific age ranges (**B09001**); and population living in group quarters (**B26001**) can be downloaded from the U.S. Census Bureau's website as CSV files for years 2007–2018. The ACS 3-year estimates were available from 2007 through 2013 and (ACS) 1-year estimates are available thereafter until 2018. ACS 1-year and 3-year estimates of total vacant housing units can also be downloaded for these years. ACS 1-year estimates for 2018 were available in September 2019, and that release schedule is typical each year. These variables can also be extracted via the Census Bureau's API.

5.4.1.2 Step 2—Calculate the State Controls Based on the Decennial Census Data and ACS Data

The NCVS includes the population in the United States who are at least 12 years old and not living in institutionalized group quarters. This step calculates the estimates of the NCVS target population in each state on the basis of the decennial census and the ACS population estimates that are publicly available. The key elements in this step are as follows:

- **Use the 2010 Census data to estimate the ratio of the institutionalized population over the group quarters populations:** Although the ACS population estimates are more up to date than the decennial census estimates, they do not distinguish the institutionalized population (which is ineligible for the NCVS) from the rest of the population, and thus the decennial census estimates are used to fill this gap. The ratio of the 2010 Census institutionalized population over the 2010 Census group quarters populations is calculated based on the 2010 Census data.
- **Use the 1-year and 5-year ACS population estimates to calculate the 3-year population estimates:** Another caveat when using the ACS population estimates is that the ACS produced 3-year estimates until 2013 but only 1-year and 5-year estimates afterward. Because the NCVS SAE procedure generates overlapping 3-year averages across a 15-year time period, the 3-year estimates of ACS are more desirable in the benchmarking procedure. To get the 3-year estimates after 2013, the average of the 1-year ACS estimates among 3 years is calculated and used as the 3-year estimates. For example, the 1-year estimates for 2015, 2016, and 2017 are averaged to create a 3-year estimate for 2017.
- **Calculate the 3-year estimates for the NCVS target population at the state level:** The following expression, which is rounded to the nearest whole number, is used to obtain the 3-year estimates for the NCVS target population at the state level:

$$ACS_{3yr} - f_{ins} \times ACS_{gq,3yr} - ACS_{age11-,3yr} - (1 - f_{ins}) \times ACS_{age18-,gq,3yr}\ , \quad (5.1)$$

where ACS_{3yr} is the ACS 3-year total population estimate, $ACS_{gq,3yr}$ is the ACS 3-year group quarters population estimate, $ACS_{age11-,3yr}$ is the ACS 3-year population estimate for those under age 11, $ACS_{age18-,gq,3yr}$ is the ACS 3-year population estimate for group quarters under age 18, and f_{ins} is the ratio of the 2010 Census institutionalized population over the 2010 Census group quarters populations.

5.4.1.3 Step 3—Export Controls Data

Two separate CSV files then can be generated and exported, one containing the estimated number of persons aged 12 and older not living in institutionalized group quarters (calculated based on equation (5.1)) and another with the 3-year estimated number of non-vacant housing units, which is the population of interest in the NCVS. The CSV files have a record for each state and a column for the estimates for each year from 2007 to 2018.

CHAPTER 6. ESTIMATION PROCEDURES TO GENERATE THE SMALL AREA ESTIMATES

6.1 Introduction

This chapter presents details on the critical R functions to fit the small area models and obtain the small area estimates. The theoretical and technical details of the dynamic models, which are the modeling methods used in the NCVS small area estimation (SAE) project, are discussed in Chapter 2. The R *state_model* function is created to fit the univariate or multivariate dynamic models for the state-level estimation, and the *substate_model* function is created to fit the univariate or multivariate dynamic models for the substate-level estimation. These functions will incorporate the results from the previous three chapters to obtain the small area estimates, including (a) the estimated covariance matrix in Chapter 3, (b) the selected predictors in the Uniform Crime Reporting (UCR) data to be used in fitting the dynamic models for the corresponding outcome variables in Chapter 4, and (c) the datasets with the UCR Summary Reporting System (SRS) estimates and the NCVS data files as produced from the data processing procedures described in Chapter 5. Note that the state-level estimates produced from the *state_model* function will go through a final benchmarking procedure as described in Chapter 7 to derive the final small area estimates. In this chapter, the state-level function is illustrated in Section 6.2, and then the substate-level function is discussed in Section 6.3. Sections 6.4 and 6.5 illustrate R program examples that use the *state_model* and *substate_model* functions for state- and county-level estimation, respectively.

Summary of Contents in Chapter 6

	Summary
What is the main point of this chapter?	This chapter describes the critical R functions to fit the small area models and obtain the small area estimates.
Why is it important?	The R functions are the workhorse of the software used in fitting the univariate or multivariate dynamic models that lead to the small area estimates.
How is it operationalized?	The R function for the state-level SAE is called *state_model* and for substate-level SAE is called *substate_model*. These functions will use the NCVS datasets and UCR SRS datasets created under procedures in Chapter 5 to generate the small area estimates of both victimization rates and prevalence rates of various types.

6.2 The *state_model* Function for State-Level Estimation

The function *state_model* is the workhorse of the software used in fitting the univariate or multivariate dynamic models leading to the small area estimates at the state level. This section will discuss the arguments and outputs of this function and provide examples of using this function to perform small area modeling to estimate both victimization and prevalence rates at the state level.

6.2.1 Arguments of the *state_model* Function

The *state_model* function has many arguments, each of which is discussed in Table 6.1. Many of these arguments have defaults that users are unlikely to need to change if they are using the data processing steps in Chapter 5, as variable names will match.

Table 6.1 *state_model* Function Arguments

Argument	Description
formula	An R formula or list of formulae defining the model
numerators	Numerator of the estimates to be fit. Defaults to being the left-hand side of the model as defined in the "formula" argument. Otherwise, it is a variable on the input dataset.
denominators	Denominator of the estimates to be fitted. Defaults to "WGTPERCY," which is the person-level analysis weight variable in NCVS. If estimating a rate (mean), one should use WGTPERCY for person-level analysis and WGTHHCY for household-level analysis with the NCVS data. Together the numerator and denominator define the weighted estimate the analyst wishes to estimate. For example: ▪ If it is specified in this function that numerators="rape" and denominators="WGTPERCY" (at person level for violent crimes), the resulting weighted estimate will be the victimization rate of rape. ▪ If it is specified in this function that numerators="burglary" and denominators="WGTHHCY" (at household level for property crimes), the resulting weighted estimate will be the victimization rate of robbery.
geocode	Geographical area level to be fitted. Defaults to "UCF_FIPSST"; defines the variable referring to the state variable on the input dataset.
sr.nsr.flag	SR and NSR indicator (self-representing or not). Defaults to "UCF_SPSUTYPE"; defines the variable referring to the SR/NSR indicator on the input dataset.

(continued)

Table 6.1 *state_model* Function Arguments (continued)

<table>
<tr><th>Argument</th><th>Description</th></tr>
<tr><td>UCR</td><td>

List of all the datasets containing the Uniform Crime Reporting (UCR) Summary Reporting System (SRS) state-level estimates. These datasets are the output files from the data processing procedures for the UCR SRS data as described in Section 5.2.2.

In the R program, one can read in the datasets with the UCR SRS estimates in two steps. First, using the following R script:

```
#specify the path to the folder that contains the UCR data
setwd(UCRdata.dir)
#Read in each dataset
Rape <- read.csv(file="Rape.legacy.est.2018.csv",
      stringsAsFactors=FALSE)
Aggravated.assault <-
read.csv(file="Aggravated.assault.est.2018.csv",
      stringsAsFactors=FALSE)
Robbery <- read.csv(file="Robbery.est.2018.csv",
      stringsAsFactors=FALSE)
Burglary <- read.csv(file="Burglary.est.2018.csv",
      stringsAsFactors=FALSE)
Motor.vehicle <- read.csv(file="Motor.vehicle.est.2018.csv",
      stringsAsFactors=FALSE)
Larceny <- read.csv(file="Larceny.est.2018.csv",
      stringsAsFactors=FALSE)
Population <- read.csv(file="Population.est.2018.csv",
      stringsAsFactors=FALSE)
```

Second, combine all the datasets in a list object named UCR:

```
UCR <- list(
  Rape=Rape, Aggravated.assault=Aggravated.assault,
  Robbery=Robbery, Burglary=Burglary,
  Motor.vehicle=Motor.vehicle, Larceny=Larceny,
  Population=Population)
```

Finally, specify `UCR=UCR` in the *state_model* function.

</td></tr>
</table>

(continued)

Table 6.1 *state_model* Function Arguments (continued)

Argument	Description

census2010	Data frame with the Census 2010 population.
	The file "apport2010_table2sorted.csv" contains the estimated population from the 2010 Census Apportionment Results (Table 2) and also provided in the supplemental files.
	In the R program, one can specify the census2010 object in the steps. First specify as follows:
	`# Read in CSV file`
	`census2010 <- read.csv(file="apport2010_table2sorted.csv",`
	`                       stringsAsFactors=FALSE)`
	`# Order the data by FIPS code`
	`census2010 <- census2010[order(census2010$statenum), ]`
	Second, specify `census2010=census2010` in the *state_model* function.

modeled.design	Data frame that contains the population size and SR population size by state according to the NCVS sampling design.
	The data file "modeled_design.csv" is provided in the supplemental files. In the R program, one can specify the modeled_design object in the steps. First, specify
	`# Read in CSV file`
	`modeled.design <- read.csv(file="modeled_design.csv",`
	`                           stringsAsFactors=FALSE)`
	`# Add on numeric FIPS code`
	`modeled.design$statenum <-`
	`  census2010$statenum[match(modeled.design$stabbr,`
	`                            census2010$stabbr)]`
	`# Order the data by FIPS code`
	`modeled.design <-`
	`  modeled.design[order(modeled.design$statenum), ]`
	
	`# Calculate the NSR pop in each state`
	`modeled.design <- within(modeled.design,`
	`                         nsrpop <- pop2010 - srpop)`
	Second, specify `modeled_design=modeled_design` in the *state_model* function.

MT	The number of areas; defaults to 51 as the number of states and the District of Columbia.

(continued)

Table 6.1 *state_model* Function Arguments (continued)

Argument	Description

file.list	List of the NCVS data files used in the SAE modeling. The creation of the NCVS data files is discussed in Chapter 5.
	To create the person-level NCVS data, one can specify an object as follows:
	`file.list.per <-`
	`  c("../Data/sastoc_per9899.Rdata",`
	`    "../Data/sastoc_per0002.Rdata",`
	`    "../Data/sastoc_per0304.Rdata",`
	`    "../Data/sastoc_per0507.Rdata",`
	`    "../Data/sastoc_per0810.Rdata",`
	`    "../Data/sastoc_per1113.Rdata",`
	`    "../Data/sastoc_per1416.Rdata",`
	`    "../Data/sastoc_per1718.Rdata")`
	To create the household-level NCVS data, one can specify an object as follows: First, specify
	`file.list.prop <-`
	`  c("../Data/sastoc_prop9899.Rdata",`
	`    "../Data/sastoc_prop0002.Rdata",`
	`    "../Data/sastoc_prop0304.Rdata",`
	`    "../Data/sastoc_prop0507.Rdata",`
	`    "../Data/sastoc_prop0810.Rdata",`
	`    "../Data/sastoc_prop1113.Rdata",`
	`    "../Data/sastoc_prop1416.Rdata",`
	`"../Data/sastoc_prop1718.Rdata")`
	Second, specify file.list=file.list.per for estimation of person-level crimes and file.list=file.list.prop for estimation of household-level crimes.

suff	Suffix of the data files. Defaults to “per.” It can be specified as either “per” or “prop” for person-level or household-level crime, respectively.
patch	An optional set of code to define other variables on the NCVS dataset. It will be discussed in Section 6.3 as a method to expand the existing code.
designvars	The NCVS pseudo-stratum and half-sample variables (see Section 3.3.3 for details about the design variables). Defaults to c("V2117", "V2118"). These are the design variables for the NCVS to estimate standard errors.
MAXITER	The maximum number of iterations to attempt. Defaults to 100. An argument to the SAE functions to control the maximum number of iterations and stops at this iteration if convergence is not reached.
PRECISION	The precision level to reach in the SAE modeling process. Defaults to .1e-4. An argument for the SAE functions to define when convergence is reached.

(continued)

Table 6.1 *state_model* Function Arguments (continued)

Argument	Description
detail	Whether output includes the covariance matrix or not; defaults to FALSE
pred.start	The year prediction will begin
pred.end	The year prediction will end
year.list	List of years to be fitted. It is not necessary that pred.start and pred.end be specified. This argument is useful when nonconsecutive years of data are analyzed. For example, one can specify year.list=list(2007, 2009, 2011, 2013). It is assumed to be in increasing order.
model	Type of model to use in the SAE modeling process. Defaults to "dyn." Specify either "dyn" or "RY" to determine type of model used (Dynamic or Rao-Yu).

Note. SAE, small area estimation.

6.2.2 Outputs from the *state_model* Function

The output from the function includes many items, even when detail=FALSE (recall that this is an argument to indicate whether the output should include the covariance matrix or not). All the items are included in the output from the *state_model* function as a list. It includes the same output as the *eblupDyn* (for the dynamic models) or *eblupRY* (for the Rao-Yu models) function in the R "sae2" package, which is archived at https://cran.r-project.org/src/contrib/Archive/sae2/. The term "eblup" stands for "empirical best linear unbiased prediction." Some examples of using the *eblupDyn* function are also illustrated in Section 2.4. The key items of the outputs from the *state_model* function are described in Table 6.2.

Table 6.2 Key Items of Output From *state_model* Function

Output	Description
eblup	The eblup estimates before applying contrasts. In the univariate case, this item contains a vector of length MT*T with the small area estimates. In the multivariate case, this item contains a data frame of MT*T rows and NV columns with the small area estimates, where T is the length of the prediction period and NV is the number of dependent variables.
eblup.mse	Mean squared error (MSE) estimates for eblup; same structure as eblup

(continued)

Table 6.2 Key Items of Output from *state_model* Function (continued)

<table>
<tr><th>Output</th><th>Description</th></tr>
<tr><td>fit</td><td>

A list summarizing the fit of the model including

- model (dynamic or Rao-Yu)
- convergence (logical)
- iterations—number of iterations
- estcoef— data frame with estimated coefficients, standard errors, *t* statistics, and p-values
- estvarcomp—data frame with estimated values of the variances and correlation coefficients and their standard errors
- goodness—the log-likelihood or restricted log-likelihood value

```
Example of a fit list:
$model
[1] "T: Dynamic, REML"

$convergence
[1] TRUE

$estcoef
                      beta   std.error     tvalue       pvalue
ucr.burglary   0.006062051 0.010010537  0.6055670 5.448023e-01
ucr.larceny    0.004296339 0.004501145  0.9544991 3.398311e-01
ucr.auto.theft 0.013413506 0.009607246  1.3961865 1.626584e-01
year01         0.094194349 0.009006562 10.4584134 0.000000e+00
year02         0.092091674 0.008858928 10.3953522 0.000000e+00
year03         0.094376601 0.008733694 10.8060353 0.000000e+00
year04         0.089673563 0.008566259 10.4682298 0.000000e+00
year05         0.085910015 0.008349608 10.2891072 0.000000e+00
year06st1      0.128640439 0.015771343  8.1565936 4.440892e-16
…
year14         0.089381558 0.006869127 13.0120692 0.000000e+00
year15         0.069085078 0.006700131 10.3110039 0.000000e+00

$estvarcomp
           estimate    std.error
sig2_u 1.960465e-09 1.638730e-06
sig2_v 2.560238e-04 7.215375e-05
rho    9.910634e-01 1.056170e-02

$iterations
```

</td></tr>
</table>

<table>
<tr><th>Output</th><th>Description</th></tr>
<tr><td></td><td><pre>
num.iter
      26

$goodness
          loglike restrictedloglike
         2205.428   1987.269
</pre></td></tr>
<tr><td>ids</td><td>A data frame with the ids (states) to identify order of estimates<pre>
             state_name pop2010 statenum stabbr
31   Alabama 4779736   1   AL
47   Alaska   710231   2   AK
39   Arizona 6392017   4   AZ
35   Arkansas 2915918   5   AR
48   California 37253956   6   CA
40   Colorado 5029196   8   CO
1   Connecticut 3574097   9   CT
22   Delaware   897934   10   DE
23 District of Columbia   601723   11   DC
24   Florida 18801310   12   FL
25   Georgia 9687653   13   GA
49   Hawaii 1360301   15   HI
41   Idaho 1567582   16   ID
10   Illinois 12830632   17   IL
11   Indiana 6483802   18   IN
15   Iowa 3046355   19   IA
16   Kansas 2853118   20   KS
32   Kentucky 4339367   21   KY
36   Louisiana 4533372   22   LA
2   Maine 1328361   23   ME
26   Maryland 5773552   24   MD
3   Massachusetts 6547629   25   MA
12   Michigan 9883640   26   MI
17   Minnesota 5303925   27   MN
33   Mississippi 2967297   28   MS
18   Missouri 5988927   29   MO
42   Montana   989415   30   MT
19   Nebraska 1826341   31   NE
43   Nevada 2700551   32   NV
4   New Hampshire 1316470   33   NH
7   New Jersey 8791894   34   NJ
44   New Mexico 2059179   35   NM
8   New York 19378102   36   NY
27   North Carolina 9535483   37   NC
</pre></td></tr>
</table>

Output	Description
	20 North Dakota 672591 38 ND 13 Ohio 11536504 39 OH 37 Oklahoma 3751351 40 OK 50 Oregon 3831074 41 OR 9 Pennsylvania 12702379 42 PA 5 Rhode Island 1052567 44 RI 28 South Carolina 4625364 45 SC 21 South Dakota 814180 46 SD 34 Tennessee 6346105 47 TN 38 Texas 25145561 48 TX 45 Utah 2763885 49 UT 6 Vermont 625741 50 VT 29 Virginia 8001024 51 VA 51 Washington 6724540 53 WA 30 West Virginia 1852994 54 WV 14 Wisconsin 5686986 55 WI 46 Wyoming 563626 56 WY
contrast.est	Estimates requested—a matrix with MT rows and NV*T columns. Contrasts are specified using the function *sae_contrasts*, which is included in the supplemental files.
contrast.mse	MSE estimates for contrast.est—a matrix with MT rows and NV*T columns

Using the function *tibble::glimpse* , a snapshot of an R script and corresponding output without details requested are shown below. This example was made using public data and random state assignments not for interpretation, but simply to provide an example of the output format.

```
prop.prev.est.nodet <- state_model(

List of 26
 $ eblup             : num [1:765] 0.1071 0.1061 0.1087 0.1039 0.0994 ...
 $ fit               :List of 6
  ..$ model      : chr "T: Dynamic, REML"
  ..$ convergence: logi TRUE
  ..$ estcoef    :'data.frame':     68 obs. of  4 variables:
  .. ..$ beta     : num [1:68] 0.00606 0.0043 0.01341 0.09419 0.09209 ...
  .. ..$ std.error: num [1:68] 0.01001 0.0045 0.00961 0.00901 0.00886 ...
  .. ..$ tvalue   : num [1:68] 0.606 0.954 1.396 10.458 10.395 ...
  .. ..$ pvalue   : num [1:68] 0.545 0.34 0.163 0 0 ...
  ..$ estvarcomp :'data.frame':     3 obs. of  2 variables:
  .. ..$ estimate : num [1:3] 1.96e-09 2.56e-04 9.91e-01
  .. ..$ std.error: num [1:3] 1.64e-06 7.22e-05 1.06e-02
  ..$ iterations : Named num 26
  .. ..- attr(*, "names")= chr "num.iter"
  ..$ goodness   : Named num [1:2] 2205 1987
  .. ..- attr(*, "names")= chr [1:2] "loglike" "restrictedloglike"
 $ parm              : Named num [1:6] 9.91e-01 1.96e-09 2.56e-04 2.21e+03 1.99e+03 ...
  ..- attr(*, "names")= chr [1:6] "rho" "sig2_u" "sig2_v" "loglikelihood" ...
 $ coef              : Named num [1:68] 0.00606 0.0043 0.01341 0.09419 0.09209 ...
  ..- attr(*, "names")= chr [1:68] "ucr.burglary" "ucr.larceny" "ucr.auto.theft" "year01" ...
 $ ids               :'data.frame': 51 obs. of  4 variables:
  ..$ state_name: chr [1:51] "Alabama" "Alaska" "Arizona" "Arkansas" ...
  ..$ pop2010   : int [1:51] 4779736 710231 6392017 2915918 37253956 5029196 3574097 897934
601723 18801310 ...
  ..$ statenum  : int [1:51] 1 2 4 5 6 8 9 10 11 12 ...
  ..$ stabbr    : chr [1:51] "AL" "AK" "AZ" "AR" ...
 $ delta             : Named num [1:3] 1.96e-09 2.56e-04 9.91e-01
  ..- attr(*, "names")= chr [1:3] "sig2_u" "sig2_v" "rho"
 $ eblup.mse         : num [1:765] 3.91e-05 3.71e-05 3.57e-05 3.51e-05 3.48e-05 ...
 $ eblup.g1          : num [1:765] 3.22e-05 3.16e-05 3.11e-05 3.05e-05 3.00e-05 ...
 $ eblup.g2          : num [1:765] 3.75e-06 3.02e-06 2.76e-06 3.23e-06 3.86e-06 ...
 $ eblup.g3          : num [1:765] 1.57e-06 1.23e-06 9.23e-07 6.84e-07 5.01e-07 ...
 $ est.fixed         : num [1:765] 0.115 0.114 0.116 0.112 0.107 ...
 $ est.fixed.var     : num [1:765] 1.38e-05 1.37e-05 1.34e-05 1.53e-05 1.53e-05 ...
 $ eblup.wt1         : num [1:765] 0.0778 0.0949 0.0898 0.0997 0.089 ...
 $ eblup.wt2         : num [1:765] 0.0976 0.1137 0.1066 0.1154 0.1296 ...
 $ contrast.est      : num [1:51, 1:15] 0.107 0.127 0.149 0.11 0.145 ...
 $ contrast.mse      : num [1:51, 1:15] 3.59e-05 1.33e-04 3.41e-05 5.33e-05 1.32e-05 ...
 $ contrast.g1       : num [1:51, 1:15] 3.16e-05 1.23e-04 2.44e-05 4.75e-05 4.56e-06 ...
 $ contrast.g2       : num [1:51, 1:15] 1.83e-06 5.07e-06 6.96e-06 3.16e-06 1.19e-06 ...
 $ contrast.g3       : num [1:51, 1:15] 1.21e-06 2.52e-06 1.37e-06 1.29e-06 3.73e-06 ...
 $ contrast.fixed.est: num [1:51, 1:15] 0.115 0.114 0.129 0.113 0.115 ...
 $ contrast.fixed.var: num [1:51, 1:15] 1.23e-05 1.15e-05 3.68e-05 1.20e-05 1.81e-05 ...
 $ contrast.wt1      : num [1:51, 1:15] 0.263 0.152 0.268 0.241 0.292 ...
 $ contrast.wt2      : num [1:51, 1:15] 0.284 0.158 0.342 0.274 0.409 ...
 $ inf.mat           : num [1:3, 1:3] 4.13e+11 9.85e+08 2.22e+07 9.85e+08 2.42e+08 ...
 $ var.coef          : num [1:68, 1:68] 1.00e-04 -2.51e-05 -1.99e-05 -1.76e-06 -3.14e-06 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:68] "ucr.burglary" "ucr.larceny" "ucr.auto.theft" "year01" ...
  .. ..$ : chr [1:68] "ucr.burglary" "ucr.larceny" "ucr.auto.theft" "year01" ...
 $ model                :Class 'formula'  language prop.prev ~ ucr.burglary + ucr.larceny +
ucr.auto.theft + year + 0
  .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
```

6.2.3 Examples of Inputs Using the *state_model* Function

The *state_model* function can be used for multiple types of estimates (victimization or prevalence rates) and univariate or multivariate estimates. The examples below illustrate these variations.

6.2.3.1 Example 1: Univariate Modeling of Prevalence of Property Crime

The following R code can estimate the state-level prevalence rate of property crime for the span of years from 2003 through 2017. In this model, the outcome variable prop.prev (prevalence proportion of property crime) is fitted by the following predictors: the UCR burglary rate, the UCR larceny rate, the UCR auto theft rate,[25] and the year. Because property is a household-level crime, the weight specified is WGTHHCY and the suff is prop. The census2010 and modeled.design are specified as well as the list of NCVS data in file.list.prop.

```
prop.prev.est <- state_model(
  prop.prev ~ ucr.burglary + ucr.larceny + ucr.auto.theft + year + 0,
  denominators="WGTHHCY",
  pred.start=2003, pred.end=2017,
  UCR=UCR, suff="prop",
  census2010=census2010, modeled.design=modeled.design,
  file.list=file.list.prop, detail=TRUE)
```

6.2.3.2 Example 2: Multivariate Modeling of Intimate Partner Violence and Other Offender Violence Victimization Rates

Two outcomes can be jointly predicted, as shown in this example, where both intimate partner violence (ipv) and other violence (otv—i.e., by someone other than an intimate partner, relative, or stranger) are jointly predicted. These are person-level crimes, so the suff=per is specified. Note that the weight is not specified because the default weight is WGTPERCY, which is appropriate in this analysis.

```
ipv.otv.est <- state_model(
  list(
    ipv ~ ucr.rape + year + 0,
    otv ~ ucr.rape + ucr.robbery + year + 0
    ),
  pred.start=2003, pred.end=2017,
  UCR=UCR, suff="per",
  census2010=census2010, modeled.design=modeled.design,
  file.list=file.list.per)
```

6.2.4 Extending to Other Estimates

The NCVS datasets created via the data processing programs discussed in Chapter 5 cover many NCVS outcomes, including the following:

[25] Please see Chapter 4 for how the three UCR variables are selected to predict property crime rates.

- Rape
- Robbery
- Simple assault
- Aggravated assault
- Personal theft
- Intimate partner violence
- Stranger violence
- All violent crime for personal crimes
- Burglary/trespassing
- Motor vehicle theft
- Other theft
- All property crime for household crimes

Although this list is long, it does not include all NCVS outcomes that an analyst may be interested in. There are two methods to extend estimates to other outcomes that are not included in the NCVS datasets created using the data processing programs discussed in Chapter 5:

- Method 1: using a patch of code to update the data in the *state_model* function
- Method 2: updating the data processing code

As an example, consider the outcome of total violent crime. This can be defined as the sum of rape, aggravated assault, simple assault, and robbery, each of which exists on the file. An analyst can write an R script defining this relationship and save to a file and then refer to the patch in the code as illustrated below.

6.2.4.1 Method 1

To implement Method 1 (using a patch of code), first insert the following code snippet in the R program before calling the *state_model* function. This code creates the two new variables, comb.asslt (combined assault) and total.viol (total violent), and saves them to a file named "*state_model_patch.R*."

```
zz <- file("state_model_patch.R", "w")
cat(" comb.asslt <- rape + aggr.asslt \n",
 " total.viol <- comb.asslt + sim.asslt + robbery \n",
 file=zz)
close(zz)
```

Second, the resulting file can be used in the patch argument under the *state_model* function to fit a model of total violent crime using predictors of the ucr.rape, the ucr.robbery, and the year.

```
tot.viol.2018 <- state_model(
  total.viol ~ ucr.rape + ucr.robbery + year + 0,
  pred.start=2003, pred.end=2017,
  UCR=UCR, suff="per",
  census2010=census2010, modeled.design=modeled.design,
  patch="state_model_patch.R",
  file.list=file.list.per, detail=TRUE)
```

6.2.4.2 *Method 2*

Alternatively, an analyst can go back to the data processing procedures and create the variable total.viol there before the estimation procedures.

6.2.4.3 *A Special Note for Prevalence Estimation*

Analysts should be mindful that, while victimization indicators can be summed as above, prevalence indicators cannot be summed. Analysts should instead use the *pmax* function to find the maximum across variables as illustrated in the following patch, where comb.asslt.prev is defined as the maximum of rape prevalence and aggravated assault prevalence and total violent prevalence is defined as the maximum of combined assault, simple assault, and robbery prevalence.

```
zz <- file("state_model_patch.R", "w")
cat(" comb.asslt.prev <- pmax(rape.prev, aggr.asslt.prev) \n",
    " total.viol.prev <-
        pmax(comb.asslt.prev, sim.asslt.prev, robbery.prev) \n",
    file=zz)
close(zz)
```

6.3 The *substate_model* Function for Substate-Level Estimation

The *substate_model* function is very similar to the *state_model* function with a few modifications, as follows:

- The geocode, sr.nsr.flag, census2010, modeled.design, MT, and year.list arguments are not used.
- The geolist argument is used and is a sorted list of FIPS codes of areas to analyze (e.g., County FIPS or CBSA FIPS).
- The state argument is used and defaults to “UCF_FIPSST.” It is used to identify the state variable the substate area is in.
- The geocode.drop argument is used and identifies areas not to be used in prediction.
- The ids argument is used and identifies the id variables.

- All inputs for the substate_model are detailed in Table 6.3. The output structure of the data coming from the substate_model has the same elements as the *state_model* function.

Table 6.3 *substate_model* Function Arguments

Argument	Description
formula	An R formula or list of formulae defining the model
numerators	Numerator of the estimates to be fit. Defaults to being the left-hand side of the model as defined in the “formula” argument. Otherwise, it is a variable on the input dataset.
denominators	Denominator of the estimates to be fitted. Defaults to “WGTPERCY,” which is the person-level analysis weight variable in NCVS. If estimating a rate (mean), one should use WGTPERCY for person-level analysis and WGTHHCY for household-level analysis with the NCVS data. Together the numerator and denominator define the weighted estimate the analyst wishes to estimate. For example: ▪ If it is specified in this function that `numerators="rape"` and `denominators="WGTPERCY"` (at person level for violent crimes), the resulting weighted estimate will be the victimization rate of rape. ▪ If it is specified in this function that `numerators= "burglary"` and `denominators= "WGTHHCY"` (at household level for property crimes), the resulting weighted estimate will be the victimization rate of robbery.

UCR	List of all the datasets containing the Uniform Crime Reporting (UCR) Summary Reporting System (SRS) state-level estimates. These datasets are the output files from the data processing procedures for the UCR SRS data as described in Section 5.2.2.
	In the R program, one can read in the datasets with the UCR SRS estimates in three steps. First, use the following R script:
	`#specify the path to the folder that contains the UCR data`
	`setwd(UCRdata.dir)`
	`Rape <- read.csv(file="Rape.legacy.co.csv",`
	`      stringsAsFactors=FALSE)`
	`Rape$yr17 <- Rape$yr16`
	`Rape$yr18 <- Rape$yr16`
	`Aggravated.assault <-`
	`read.csv(file="Aggravated.assault.co.csv",`
	`      stringsAsFactors=FALSE)`
	`Aggravated.assault$yr17 <- Aggravated.assault$yr16`
	`Aggravated.assault$yr18 <- Aggravated.assault$yr16`
	`Robbery <- read.csv(file="Robbery.co.csv",`
	`      stringsAsFactors=FALSE)`
	`Robbery$yr17 <- Robbery$yr16`
	`Robbery$yr18 <- Robbery$yr16`
	`Burglary <- read.csv(file="Burglary.co.csv",`
	`      stringsAsFactors=FALSE)`
	`Burglary$yr17 <- Burglary$yr16`
	`Burglary$yr18 <- Burglary$yr16`

Argument	Description

	`Motor.vehicle <- read.csv(file="Motor.vehicle.co.csv",`
	`      stringsAsFactors=FALSE)`
	`Motor.vehicle$yr17 <- Motor.vehicle$yr16`
	`Motor.vehicle$yr18 <- Motor.vehicle$yr16`
	`Larceny <- read.csv(file="Larceny.co.csv",`
	`      stringsAsFactors=FALSE)`
	`Larceny$yr17 <- Larceny$yr16`
	`Larceny$yr18 <- Larceny$yr16`
	Second, combine all the datasets in a list object named UCR:
	`UCR <- list(`
	`      Rape=Rape, Aggravated.assault=Aggravated.assault,`
	`      Robbery=Robbery, Burglary=Burglary,`
	`      Motor.vehicle=Motor.vehicle, Larceny=Larceny)`
	Third, specify `UCR=UCR` in the *substate_model* function.

file.list	List of the NCVS data files used in the SAE modeling. The creation of the NCVS data files is discussed in Chapter 5.
	To create the person-level NCVS data, one can specify an object as follows:
	`file.list.per <-`
	`  c("../Data/sastoc_per9899.Rdata",`
	`    "../Data/sastoc_per0002.Rdata",`
	`    "../Data/sastoc_per0304.Rdata",`
	`    "../Data/sastoc_per0507.Rdata",`
	`    "../Data/sastoc_per0810.Rdata",`
	`    "../Data/sastoc_per1113.Rdata",`
	`    "../Data/sastoc_per1416.Rdata",`
	`    "../Data/sastoc_per1718.Rdata")`
	To create the household-level NCVS data, one can specify an object as follows:
	`file.list.prop <-`
	`  c("../Data/sastoc_prop9899.Rdata",`
	`    "../Data/sastoc_prop0002.Rdata",`
	`    "../Data/sastoc_prop0304.Rdata",`
	`    "../Data/sastoc_prop0507.Rdata",`
	`    "../Data/sastoc_prop0810.Rdata",`
	`    "../Data/sastoc_prop1113.Rdata",`
	`    "../Data/sastoc_prop1416.Rdata",`
	`    "../Data/sastoc_prop1718.Rdata")`
	Then, specify file.list=file.list.per for estimation of person-level crimes and file.list=file.list.prop for estimation of household-level crimes.

state	Defines the variable referring to the state variable on the input dataset; defaults to “UCF_FIPSST”
suff	Suffix of the data files. Defaults to “per.” It can be specified as either “per” or “prop” for person-level or household-level crime, respectively.
patch	An optional set of code to define other variables on the NCVS dataset. It will be discussed in Section 6.3 as a method to expand the existing code.

(continued)

Table 6.3 *substate_model* Function Arguments (continued)

Argument	Description
designvars	The NCVS pseudo-stratum and half-sample variables (see Section 3.3.3 for details about the design variables). Defaults to c("V2117", "V2118"). These are the design variables for the NCVS to estimate standard errors.
MAXITER	The maximum number of iterations to attempt. Defaults to 100. An argument to the SAE functions to control the maximum number of iterations and stops at this iteration if convergence is not reached.
PRECISION	The precision level to reach in the SAE modeling process. Defaults to .1e-4. An argument for the SAE functions to define when convergence is reached.
detail	Whether output includes the covariance matrix or not; defaults to FALSE
pred.start	The year prediction will begin
pred.end	The year prediction will end
model	Type of model to use in the SAE modeling process. Defaults to “dyn.” Specify either “dyn” or “RY” to determine type of model used (dynamic or Rao-Yu).
geocode.drop	FIPS codes of areas not to use in modeling. Predictions are still produced for this error. This might be used if there is poor data quality in the UCR data.
ids	A data frame with the FIPS codes of areas to be predicted and any other identifying information to save on predictions—e.g., county name, state name

Note. SAE, small area estimation.

6.4 SAE Program for State-Level Model

In addition to using the function *state_model*, analysts must read in the necessary functions and data files in the R script to implement the SAE procedure. An example R script (named *FitsExamples.R*) is included in the supplemental files along with the R function files called within this R script. This example R script includes examples of prevalence and victimization rate estimation for both violent and property crimes at the state level. This section provides detailed explanation for this example script by using the prevalence rate estimation of property crime as the example.

To implement the state-level SAE program as demonstrated in the example R script, the following steps are necessary.

Step 1. Declare the paths of folders that contain the required R function files and datasets to be used in the SAE procedure. The folders are as follows:

a. **program.dir:** This folder contains all the critical R function files to be called in the main SAE program, including (1) the R script file "*loadnewNCVSsae2package.R*," (2) the R script file "*loadnewsae2package.R*," (3) the folder "newNCVSsae2" and (4) the folder "newsae2." The last two folders contain a series of R function files. All the R script files and folders can be found in the supplemental files.

b. **UCRdata.dir:** This folder contains the output data files of UCR SRS state-level estimates generated based on the data processing procedures described in Chapter 5. Variables from the UCR SRS data will be used as predictors in the SAE models.

c. **newproduction.dir:** This is the folder where the apport2010_table2sorted.csv and modeled_design.csv are saved. These files can be found in the supplemental files. These two files will be used in the *state_model* function as shown in Step 7.

d. **results.dir:** This is the folder where the small area estimate results should be saved.

```
program.dir <- "" #folder with SAE code with subfolders of newNCVSsae2
and newsae2
UCRdata.dir    <- "" #folder with UCR data
newproduction.dir <- "" #folder with apport2010_table2sorted.csv and
modeled_design.csv
results.dir <- "" #folder to save results
```

Step 2. Load the R codes for SAE in "*loadnewNCVSsae2package.R*" and "*loadnewsae2package.R*." The "*loadnewNCVSsae2package.R*" file will then load all the R function files in the "newNCVSsae2" folder, and the "*loadnewsae2package.R*" file will load all the R function files in the "newsae2" folder.

```
setwd(program.dir)
source("loadnewsae2package.R")
source("loadnewNCVSsae2package.R")
```

Step 3. Create two lists of NCVS data file locations, one for person-level NCVS data (file.list.per) and one for household-level NCVS data (file.list.prop). The NCVS data files are the output files generated based on the data processing procedures for NCVS data as described in Chapter 5.

```
#Create the list of file locations for the NCVS person-level data files
file.list.per <- c("../sastoc_per9899.Rdata", "../sastoc_per0002.Rdata",
  "../sastoc_per0304.Rdata", "../sastoc_per0507.Rdata",
  "../sastoc_per0810.Rdata", "../sastoc_per1113.Rdata",
  "../sastoc_per1415.Rdata")

#Create the list of file locations for the NCVS household-level data files
file.list.prop <- c("../sastoc_prop9899.Rdata", "../sastoc_prop0002.Rdata",
  "../sastoc_prop0304.Rdata", "../sastoc_prop0507.Rdata",
  "../sastoc_prop0810.Rdata", "../sastoc_prop1113.Rdata",
  "../sastoc_prop1415.Rdata")
```

Step 4. Load the necessary R packages—survey, MASS, and readr—and set up options for the survey package.

```
library(survey)
options(survey.lonely.psu="adjust")
library(MASS)
library(readr)
```

Step 5. Set up the span of years for the analysis using pred.start (the year prediction will begin) and pred.end (the year prediction will end). The value T will be the number of years in the specified span of years. This is done globally so you can use the same specifications throughout all your analyses in one script.

```
pred.start <- 2004
pred.end   <- 2015
T <- pred.end - pred.start + 1
```

Step 6. Read in the UCR SRS data files, which are the output data files from the data processing procedures to get the state-level UCR SRS estimates as described in Chapter 5.

```
setwd(UCRdata.dir)
Rape <- read.csv(file="Rape.legacy.est.csv", stringsAsFactors=FALSE)
Aggravated.assault <- read.csv(file="Aggravated.assault.est.csv",
      stringsAsFactors=FALSE)
Robbery <- read.csv(file="Robbery.est.csv", stringsAsFactors=FALSE)
Burglary <- read.csv(file="Burglary.est.csv", stringsAsFactors=FALSE)
Motor.vehicle <- read.csv(file="Motor.vehicle.est.csv",
stringsAsFactors=FALSE)
Larceny <- read.csv(file="Larceny.est.csv", stringsAsFactors=FALSE)
Population <- read.csv(file="Population.est.csv", stringsAsFactors=FALSE)

UCR <- list(Rape=Rape, Aggravated.assault=Aggravated.assault,
Robbery=Robbery,
            Burglary=Burglary, Motor.vehicle=Motor.vehicle, Larceny=Larceny,
            Population=Population)
```

Step 7. Read in and reformat the apport2010_table2sorted.csv and modeled_design.csv files.

```
setwd(newproduction.dir)

# file of census state counts
census2010 <- read.csv(file="apport2010_table2sorted.csv",
stringsAsFactors=FALSE)
census2010 <- census2010[order(census2010$statenum), ]

# modeled proportions of self- and nonself-representing strata
modeled.design <- read.csv(file="modeled_design.csv", stringsAsFactors=FALSE)

modeled.design$statenum <- census2010$statenum[match(modeled.design$stabbr,
                                                     census2010$stabbr)]
modeled.design <- modeled.design[order(modeled.design$statenum), ]

sr.pop.us <- as.numeric(sum(modeled.design$srpop))
pop.us <-    as.numeric(sum(modeled.design$pop2010))
nsr.pop.us <- pop.us - sr.pop.us
modeled.design <- within(modeled.design, nsrpop <- pop2010 - srpop)
```

Step 8. Calculate the state-level small area estimates for prevalence rate of total property crime using the *state_model* function.

```
# Estimate the state-level prevalence rate of total property crime using the
state_level function

## Specify detail=TURE so that the output will include the covariance matrix
prop.prev.est <- state_model(
  prop.prev ~ ucr.burglary + ucr.larceny + ucr.auto.theft + year + 0,
  denominators="WGTHHCY",
  pred.start=2001, pred.end=2015,
  UCR=UCR, suff="prop",
  census2010=census2010, modeled.design=modeled.design,
  file.list=file.list.prop, detail=TRUE)

## Specify detail=FALSE so that the output will not include the covariance
matrix
prop.prev.est.nodet <- state_model(
  prop.prev ~ ucr.burglary + ucr.larceny + ucr.auto.theft + year + 0,
  denominators="WGTHHCY",
  pred.start=2001, pred.end=2015,
  UCR=UCR, suff="prop",
  census2010=census2010, modeled.design=modeled.design,
  file.list=file.list.prop, detail=FALSE)
```

Step 9. Save the small area estimates as rds files in the result folder.

```
setwd(results.dir)
write_rds(prop.prev.est, "prop.prev.est.det.rds")
write_rds(prop.prev.est.nodet, "prop.prev.est.nodet.rds")
```

6.5 SAE Program for Substate-Level Model

Similar to the state-level estimation, the R script used for SAE has several components in addition to using the function *substate_model*. Analysts must also read in the necessary functions and data files to implement the analysis. An example script is included in the supplemental files (named *FitsExamples_county.R*) along with the R function files called within this R script. This example R script includes examples of prevalence and victimization rate estimation for both violent and property crimes at the county level. This section provides detailed explanation for this example script by using the victimization rate estimation of violent crime as the example.

To implement the substate-level SAE program as demonstrated in the example R script, the following steps are necessary.

Step 1. Declare the paths of folders that contain the required R function files and datasets to be used in the SAE procedure. The folders are as follows:

a. **program.dir:** This folder contains all the critical R function files to be called in the main SAE program, including (1) the R script file "*loadnewNCVSsae2package.R*," (2) the R script file "*loadnewsae2package.R*," (3) the folder "newNCVSsae2" and (4) the folder "newsae2." The last two folders contain a series of R function files. All the R script files and folders can be found in the supplemental files.

b. **UCRdata.dir:** This folder contains the output data files of UCR SRS county-level estimates generated based on the data processing procedures described in Chapter 5. Variables from the UCR SRS data will be used as predictors in the SAE models.

c. **newproduction.dir:** This is the folder where the apport2010_table2sorted.csv and modeled_design.csv are saved. These files can be found in the supplemental files. These two files will be used in the *state_model* function as shown in Step 7.

d. **results.dir:** This is the folder where the small area estimate results should be saved.

```
program.dir <- "" #folder with SAE code with subfolders of newNCVSsae2
and newsae2
UCRdata.dir    <- "" #folder with UCR data
newproduction.dir <- "" #folder with apport2010_table2sorted.csv and
modeled_design.csv
results.dir <- "" #folder to save results
```

Step 2. Load the R codes for SAE in "*loadnewNCVSsae2package.R*" and "*loadnewsae2package.R*." The "*loadnewNCVSsae2package.R*" file will then load all the R function files in the "newNCVSsae2" folder, and the "*loadnewsae2package.R*" file will load all the R function files in the "newsae2" folder.

```
setwd(program.dir)
source("loadnewsae2package.R")
source("loadnewNCVSsae2package.R")
```

Step 3. Create two lists of NCVS data file locations, one for person-level NCVS data (file.list.per) and one for household-level NCVS data (file.list.prop). The NCVS data files are the output files generated based on the data processing procedures for NCVS data as described in Chapter 5.

```
file.list.per <- c("../sastoc_per9899.Rdata", "../sastoc_per0002.Rdata",
  "../sastoc_per0304.Rdata", "../sastoc_per0507.Rdata",
  "../sastoc_per0810.Rdata", "../sastoc_per1113.Rdata",
  "../sastoc_per1415.Rdata")

file.list.prop <- c("../sastoc_prop9899.Rdata", "../sastoc_prop0002.Rdata",
  "../sastoc_prop0304.Rdata", "../sastoc_prop0507.Rdata",
  "../sastoc_prop0810.Rdata", "../sastoc_prop1113.Rdata",
  "../sastoc_prop1415.Rdata")
```

Step 4. Load the necessary R packages—survey, MASS, and readr—and set up options for the survey package.

```
library(survey)
options(survey.lonely.psu="adjust")
library(MASS)
library(readr)
```

Step 5. Set up the span of years for the analysis using pred.start (the year prediction will begin) and pred.end (the year prediction will end). The value T will be the number of years in the specified span of years. This is done globally so you can use the same specifications throughout all your analyses in one script.

```
pred.start <- 2007
pred.end   <- 2018
T <- pred.end - pred.start + 1
```

Step 6. Read in the UCR SRS data files, which are the output data files from the data processing procedures to get the county-level UCR SRS estimates as described in Chapter 5. The 2017 and 2018 UCR SRS data are not available at the county level, so use the 2016 data to represent 2017 and 2018 as well.

```
setwd(UCRdata.dir)
# 2017 and 2018 data not available. Use 2016 as 2017 and 2018 data
Rape <- read.csv(file="Rape.legacy.co.csv",
      stringsAsFactors=FALSE)
Rape$yr17 <- Rape$yr16
Rape$yr18 <- Rape$yr16
Aggravated.assault <- read.csv(file="Aggravated.assault.co.csv",
      stringsAsFactors=FALSE)
Aggravated.assault$yr17 <- Aggravated.assault$yr16
Aggravated.assault$yr18 <- Aggravated.assault$yr16
Robbery <- read.csv(file="Robbery.co.csv",
      stringsAsFactors=FALSE)
Robbery$yr17 <- Robbery$yr16
Robbery$yr18 <- Robbery$yr16
Burglary <- read.csv(file="Burglary.co.csv",
      stringsAsFactors=FALSE)
Burglary$yr17 <- Burglary$yr16
Burglary$yr18 <- Burglary$yr16
Motor.vehicle <- read.csv(file="Motor.vehicle.co.csv",
      stringsAsFactors=FALSE)
Motor.vehicle$yr17 <- Motor.vehicle$yr16
Motor.vehicle$yr18 <- Motor.vehicle$yr16
Larceny <- read.csv(file="Larceny.co.csv",
      stringsAsFactors=FALSE)
Larceny$yr17 <- Larceny$yr16
Larceny$yr18 <- Larceny$yr16
UCR <- list(
      Rape=Rape, Aggravated.assault=Aggravated.assault,
      Robbery=Robbery, Burglary=Burglary,
      Motor.vehicle=Motor.vehicle, Larceny=Larceny)
```

Step 7. Read in county data file (county.data.csv), an example of which is included in the supplemental files. This file provides the list of counties to analyze.

```
county.data <- read.csv("county.data.csv", stringsAsFactors=FALSE)
ids <- county.data[, c("CTYSTNAME", "CENSUS2010POP", "fips",     "stabbr")]
```

Step 8. Set up a patch file that combines the New York City counties into one and updates the FIPS code for Miami-Dade County (FL), which was changed in 1997.

```
setwd(program.dir)

zz <- file("county_model_patch.R", "w")
cat("   comb.asslt <- rape + aggr.asslt \n",
    "   total.viol <- comb.asslt + sim.asslt + robbery \n",
    "   fips  <- 1000 * UCF_FIPSST + UCF_FIPSCO \n",
    "   fips[fips %in% c(36005, 36047, 36081, 36085)] <- 36061 \n",
    "   fips[fips == 12025] <- 12086 \n",
        file=zz)
close(zz)
```

Step 9. Calculate the county-level small area estimates for victimization rates of total violent crime using the *substate_model* function.

```
# model including Illinois counties

tot.viol.2018a <- substate_model(
# (1a)
      total.viol ~ ucr.rape + ucr.robbery + year + 0,
      geolist=county.data$fips, UCR=UCR,
      file.list=file.list.per, patch="county_model_patch.R",
      pred.start=pred.start, pred.end=pred.end, suff="per",
      ids=ids)

# model excluding Illinois counties from regression using the argument
geocode.drop=c(17031, 17043)

tot.viol.2018b <- substate_model(
# (1b)
      total.viol ~ ucr.rape + ucr.robbery + year + 0,
      geolist=county.data$fips, UCR=UCR,
      file.list=file.list.per, patch="county_model_patch.R",
      pred.start=pred.start, pred.end=pred.end, suff="per",
      geocode.drop=c(17031, 17043), ids=ids)

# model for 2011-2018 including Illinois counties
pred.start <- 2011
pred.end <- 2018

tot.viol.2018c <- substate_model(
# (1c)
      total.viol ~ ucr.rape + ucr.robbery + year + 0,
      geolist=county.data$fips, UCR=UCR,
      file.list=file.list.per, patch="county_model_patch.R",
      pred.start=pred.start, pred.end=pred.end, suff="per",
      ids=ids)

# model for 2007-2018 including Illinois counties using UCR SRS robbery
(ucr.robbery) as the single predictor in the model
pred.start <- 2007
pred.end <- 2018

tot.viol.2018d <- substate_model(
# (1d)
      total.viol ~ ucr.robbery + year + 0,
      geolist=county.data$fips, UCR=UCR,
      file.list=file.list.per, patch="county_model_patch.R",
      pred.start=pred.start, pred.end=pred.end, suff="per",
      ids=ids)
```

Step 10. Save the small area estimates as a *.Rdata* file in the result folder.

```
setwd(results.dir)

save(list=c("tot.viol.2018a", "tot.viol.2018b", "tot.viol.2018c",
            "tot.viol.2018d"), file="fits20Nov2019a.Rdata")
```

CHAPTER 7. BENCHMARKING PROCEDURES FOR STATE-LEVEL SMALL AREA ESTIMATES

7.1 Introduction

The state estimates produced from the main small area estimation (SAE) modeling process using the *state_model* function need go through a final benchmarking procedure to derive the final small area estimates. The benchmarking procedure will address two separate issues:

1. Benchmarking victimization or prevalence rates so that the sums of the estimates by subtypes, which are distinguishing components of a major class of crime, agree with the estimates of their aggregated types
2. Benchmarking victimization or prevalence rates and their estimated numbers at the state level so that the sums of estimated numbers are consistent with the published National Crime Victimization Survey (NCVS) national totals

The mathematical solution to address the first issue is discussed in Section 7.2 and then the developed benchmarking procedures that can address both the first and second issues are illustrated in Sections 7.3 and 7.4. Note that separate benchmarking procedures are developed for
(a) victimization rates (and the estimated numbers of victimizations) for violent crimes;
(b) victimization rates (and the estimated numbers of victimizations) for property crimes;
(c) prevalence rates (and the estimated prevalence counts) for violent crimes; and (d) prevalence rates (and the estimated prevalence counts) for property crimes. Section 7.5 provides details on software implementation for the final benchmarking procedures.

Summary of Contents in Chapter 7

What is the main point of this chapter?	This chapter describes the final benchmarking procedures (i.e., procedures to ensure that the sum of the small area estimates of subdomains agree with the small area estimates of their aggregated domains) for the state-level small area estimates.
Why is it important?	The benchmarking procedures adjust the small area estimates so that the estimates of crime subtypes agree with the estimate of their aggregated type, and the sum of the state-level estimates agrees with the published national totals in NCVS.
How is it operationalized?	The estimates of subtypes and their aggregated types will be estimated separately in the estimation procedures. The state-level estimates will then be adjusted to meet their published national totals in NCVS. The estimates of subtypes will be further adjusted so that they can agree with the estimate of their aggregated type.

7.2 Benchmarking Estimates of Crime Subtypes

7.2.1 Why Do We Need to Benchmark Estimates of Crime Subtypes?

In the NCVS SAE, estimates for the set of different subtypes of total violent (or property) crime are produced simultaneously, based on a multivariate dynamic model, whereas the estimate for the total violent crime (or property) crime is derived from a univariate dynamic model. A previous evaluation found that the univariate modeling method seemed to give more stable results than just summing up the estimates by subtypes to derive estimates for total violent or property crime—presumably because the univariate model is based on more data and estimates fewer parameters.

However, because different models are used to produce estimates for different crime subtypes and their sums, the sum of the estimates by subtypes does not necessarily agree with the estimates of their aggregated type. Therefore, a top-down benchmarking approach was developed and can be implemented for both state-level and substate-level estimates. This approach adjusts estimates for subtypes so that their sum will agree with the estimates from univariate models for their aggregated type. The implementation of this approach is slightly different between victimization rates and prevalence rates, as presented in the following subsections.

7.2.2 What Crime Subtypes Are Formed and Estimated in the NCVS SAE?

Figures 7.1 and 7.2 illustrate different subtypes of total violent crime and total property crime that are formed and estimated in the NCVS SAE. Figure 7.1 diagrams how the total violent crime is dichotomized into its two subtypes using the following steps:

- Violent crime by relationship:
  - First, total violent crime is dichotomized into violent crime by strangers and violent crime by non-strangers.
  - Second, violent crime by all non-strangers is furtherly dichotomized into intimate partner violence and violence by all other non-strangers.
- Violent crime by type:
  - First, total violent crime is dichotomized into simple assault and violent crime excluding simple assault (called “serious violent crime” in publications before 2018).

- Second, violent crime excluding assault is furtherly dichotomized into robbery and aggravated assault/rape.

Figure 7.1 Dichotomization of Total Violent Crime Into Subtypes

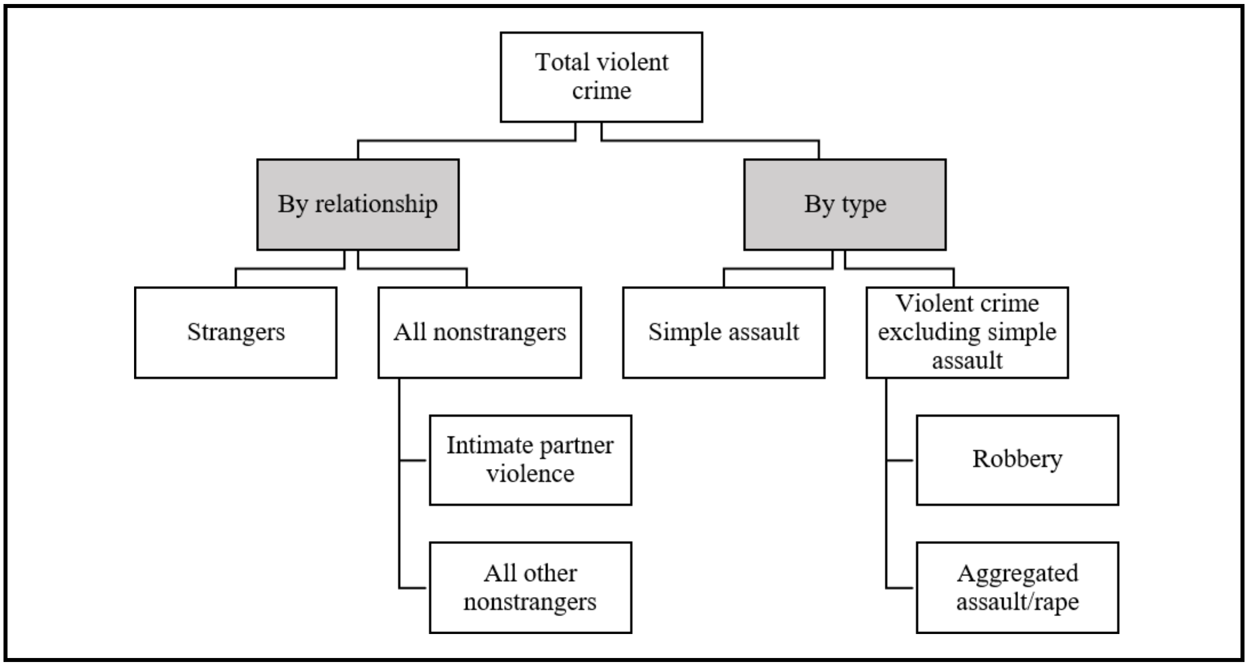
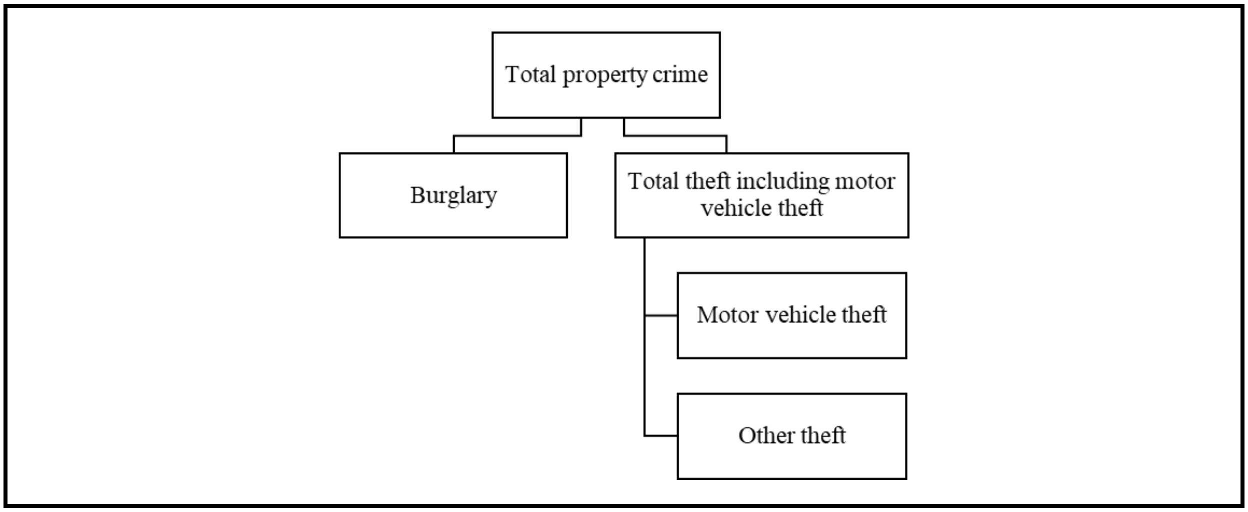


Figure 7.2 diagrams how the total property crime is dichotomized into its subtypes. First, total property crime is dichotomized into burglary and total theft including motor vehicle theft. Second, the total theft including motor vehicle theft is furtherly dichotomized into motor vehicle theft and other theft.

Figure 7.2 Dichotomization of Total Property Crime Into Subtypes



7.2.3 Top-Down Benchmarking Approach for Estimation of Victimization Rate

Suppose a total for a major class of crime (total violent or property crime), M, is broken down into three subtypes, A, B, and C. Let $\hat{p}_{M,it}$ denote the small area estimate of the victimization rate of M in state i in a given year *t* from a univariate model, and if $\hat{p}_{A,it}$, $\hat{p}_{B,it}$, and $\hat{p}_{C,it}$ are estimates of the subtypes either from a multivariate model (as used in this project) or three univariate models (which can be another option to obtain the estimates for subtypes), then the estimates of subtypes can each be multiplied by the factor $\hat{p}_{M,it}/(\hat{p}_{A,it} + \hat{p}_{B,it} + \hat{p}_{C,it})$ so that the sum of the subtypes will agree with the estimated total.

Sometimes, not only the estimates of total type M and its subtypes A, B, C are of interest, but the estimates of the combined subtypes also need to be produced. For example, as shown in Figure 7.2, the subtypes of total property crime (M) can be divided into burglary (A), motor vehicle theft (B), and other theft (C). The combined subtype of B and C, total theft including motor vehicle theft, needs also to be estimated. In this case, the model can dichotomize the total into two components first (A and B+C), modeling the two components first and then separately modeling the two components (B and C) that had been combined at the first steps. Conversely, the estimates of $\hat{p}_{A,it}$, and $\hat{p}_{BC,it}$ from the first step can be multiplied by $f_{1,it} = \hat{p}_{M,i}/(\hat{p}_{A,it} + \hat{p}_{BC,it})$ so that their sum will agree with the estimate of total type M. Then, the estimates of $\hat{p}_{B,it}$, and $\hat{p}_{C,it}$ from the second step will be multiplied by $f_{2,it} = f_{1,it}\hat{p}_{BC,i}/(\hat{p}_{B,it} + \hat{p}_{C,it})$ so that they will agree with the benchmarked estimate of the combined B and C subtype. Another advantage of modeling the subtypes through this dichotomizing procedure with multiple steps is that modeling the combined subtype is considered to give more stable results, and thus benchmarking the estimates of subtypes to the estimates of their combined types can improve the reliability of each individual estimate of those subtypes.

For the 2007–2018 state estimates, total violent crime is first dichotomized into violent crime except simple assault and simple assault, and then these two components are modeled. The next model is of the dichotomy aggravated assault/rape and robbery. Similarly, property crime is dichotomized into burglary and combined theft, and then combined theft is dichotomized into motor vehicle theft and larceny.

7.2.4 Top-Down Benchmarking for Estimation of Prevalence Rate

Benchmarking for prevalence rates is slightly different from benchmarking for victimization rates because an inequality rather than an equality is involved. That is, the sum of prevalence rates for subtypes A, B, and C can exceed the prevalence of M, but it cannot be less. This condition requires an additional initial step to estimate the sum of the prevalence rates of subtypes.

For the initial step, denote the state-level estimates of the sum of the prevalence rates of subtypes as $\hat{p}_{S,it}$. To attempt to impose the logical relationship $\hat{p}_{S,it} \geq \hat{p}_{M,it}$, the ratio $r_{it} = p_{S,it}/p_{M,it}$ is modeled via a univariate model based on observed ratios, and then the estimate $\hat{p}_{S,it} = \hat{r}_{it}\hat{p}_{M,it}$ is formed. The adjustment factor for the prevalence rates of subtypes A, B, and C is then $\hat{p}_{S,it}/(\hat{p}_{A,it} + \hat{p}_{B,it} + \hat{p}_{C,it})$.

When a total is dichotomized by combining two of the initial subtypes, the logic of benchmarking the prevalence rate is similar to the logic of benchmarking the victimization rate. If the previous example is used to illustrate this logic, the following steps are implemented. In the first step, the crime type M is dichotomized into subtypes A and BC and $\hat{p}_{A,it}$ and $\hat{p}_{BC,it}$ are produced from the SAE modeling procedure. The ratio $\hat{r}_{1,it}$ is first modeled based on the sum of $\hat{p}_{A,it}$ and $\hat{p}_{BC,it}$ and their total prevalence $\hat{p}_{M,it}$. The estimated sum of the prevalence rates of A and BC is then estimated by $\hat{p}_{1S,it} = \hat{r}_{1,it}\hat{p}_{M,it}$ and used to compute the adjustment factor

$f_{1,it} = \hat{p}_{1S,it}/(\hat{p}_{A,it} + \hat{p}_{BC,it})$ for prevalence rates $\hat{p}_{A,it}$ and $\hat{p}_{BC,it}$. In the second step, B and C are separately modeled and $\hat{p}_{B,it}$ and $\hat{p}_{C,it}$ are produced. The ratio $\hat{r}_{2,it}$ is then modeled based on the sum of $\hat{p}_{B,it}$ and $\hat{p}_{C,it}$ and their total prevalence $\hat{p}_{BC,it}$ (from the first step) and used to compute the factor $f_{2,it} = f_{1,it}\hat{r}_{2,i}t\ \hat{p}_{BC,it}/(\hat{p}_{B,it} + \hat{p}_{C,it})$ to adjust $\hat{p}_{B,it}$ and $\hat{p}_{C,it}$.

7.3 Benchmarking Victimization Rates and the Estimated Numbers of Victimizations

7.3.1 For Violent Crimes

For a certain crime type M (e.g., robbery) in a state *i* given a 3-year period *t* (e.g., 2014–2016), the numbers of victimizations for violent crimes are estimated by multiplying the small area estimates of the victimization rates (e.g., $\hat{p}_{M,it}$, $\hat{p}_{A,it}$, $\hat{p}_{B,it}$, and $\hat{p}_{C,it}$) by the NCVS population estimate in this state and 3-year period. The benchmarking method first adjusts the state-level estimated numbers of victimizations of total violent crime so that their sum agrees with the NCVS national totals, which are considered to be more stable given that the national-level sample size is much larger. Then, the state-level estimated numbers of victimizations by crime subtypes will be adjusted via a raking method so that their sum will agree with the adjusted state-level estimates of total violent crime and each estimated number will agree with the NCVS national estimates of the corresponding subtype.

Benchmarking for violent crimes is implemented through *benchmark_viol_ser_2018_revision.R* as included in the supplemental files. The following steps were used in this program to benchmark the state estimates to reach this goal:

1. **Calculate the initial numbers of victimizations for each state:** For each state and 3-year period, the initial estimated numbers of victimizations for total violent crime and by violent crime subtypes can be calculated by multiplying the population estimates by the small area estimates of their corresponding victimization rates. The population estimates are the 3-year average of the population estimates for the NCVS target population at the state level as described in Section 5.4 (Calculating State Controls for the Benchmarking Procedure Using the Census and ACS Data).
2. **Calculate the national totals:** For each 3-year period (e.g., 2014–2016), the national totals for total violent crime and by violent crime subtypes are calculated by summing the records from the NCVS personal-level data (e.g., *sastoc_per1416.Rdata*) that are generated based on information described in Section 5.3 and used as input data to produce the small area estimates for violent crimes. The benchmarking R program also displays the national totals for checking against the published values.

3. **Adjust the initial number of victimizations of total violent crime:** The initial estimates of total violent crime from Step 1 then are proportionally adjusted to agree with the NCVS 3-year average estimate of total violent crime from Step 2.

4. **Rake the initial number of victimizations to adjust by crime subtypes:** Total violent crime can be dichotomized into different sets of subtypes, as shown in Figure 7.1. The initial estimates of each subtype from Step 1 are adjusted using two-dimensional raking (iterative proportional fitting) so that the estimates agree with the NCVS national estimates of the corresponding subtype (the first margin in Step 2) and the adjusted state estimates of total violent crime from Step 3 (the second margin). These sets include the following:

   - ▪ Violent crime by relationship: Crime by strangers and all non-strangers

     – All non-strangers by type: This subtype is furtherly dichotomized into intimate partner violence and violence by all other non-strangers. Again, the initial estimates of these two subtypes are raked to their national estimates and the raking-adjusted estimates for all non-strangers.

   - ▪ Violent crime by type: Simple assault and violent crime excluding simple assault

     – Violent crime excluding simple assault: This subtype is furtherly dichotomized into robbery and aggravated assault/rape. Again, the initial estimates of these two subtypes are raked to their national estimates and the raking-adjusted estimates for violent crime excluding simple assault.

5. **Convert the adjusted estimates into victimization rates:** Adjusted estimates from Steps 3 and 4 are converted to rates per 1,000 based on the population estimates as described in Step 1 and rounded. Rounding is incorporated to meet requirements of the Census Bureau's Disclosure Review Board, both for the estimated rates and estimated totals.

6. **Convert the root mean square errors (RMSEs) into rates:** The RMSEs are also converted to rates per 1,000 and rounded, based on the results from the model used in estimating each type. Note that the RMSE estimates are not adjusted for the effect of the raking, which would require additional methodological development. For each characteristic that is benchmarked, separate CSV files are output for the rate, the RMSE of the rate, and the estimated total for years, beginning with 2005–2007.

7.3.2 For Property Crimes

Benchmarking the victimization rates and numbers for property crimes (implemented through *benchmark_prop_cen_revision.R*) is similar to benchmarking for violent crimes with the following modifications:

1, The numbers of victimizations for property crimes for each state and 3-year period are estimated by multiplying the small area estimates of the victimization rates by the 3-year averages of NCVS household estimates (instead of by the NCVS population estimates for in this state and 3-year period).

2. The national totals for total property crime and by subtypes are calculated by summing the records from the NCVS household-level data (e.g., *sastoc_prop1416.Rdata*), rather than person-level data. The household-level data are generated based on information described in Section 5.3 and used as input data to produce the small area estimates for property crimes.
3. In Step 4, the total property crime is first dichotomized into burglary and total theft including motor vehicle theft. The initial estimates of these two subtypes are adjusted through two-dimensional raking. Then, the total theft including motor vehicle theft is furtherly dichotomized into motor vehicle theft and other theft. Again, the initial estimates of these two subtypes are raked to their national estimates and the raking-adjusted estimates for total theft including motor vehicle theft.

7.4 Benchmarking Prevalence Rates and the Estimated Prevalence Counts

7.4.1 For Violent Crimes

Prevalence counts (number of victims) for each state and 3-year period are estimated by multiplying the small area estimates of the prevalence rates by the NCVS population estimate in this state and 3-year period. The principles for benchmarking prevalence rates are generally similar to those for victimization rates. However, as described in Section 7.2.4, unlike victimization rates, the sum of prevalence rates for subtypes A, B, and C can exceed the prevalence of M, but it cannot be less. Thus, the sum of the prevalence rates of subtypes must be estimated first before the small area estimates of prevalence rates of these subtypes can be benchmarked.

Benchmarking for prevalence rates of violent crimes is implemented through *benchmark_viol_prev_2018_revision.R* as included in the supplemental files. The following steps were used in this program to benchmark the state estimates to reach this goal:

1. **Calculate the initial prevalence counts for each state:** For each state and 3-year period, the initial estimated prevalence counts for total violent crime and by violent crime subtype can be calculated by multiplying the population estimates by the small area estimates of their corresponding prevalence rates.
2. **Total violent crime can be divided into different sets of subtypes (e.g., violent crime by relationship, violent crime by type).** For each set of subtypes, the initial estimate of the sum of the prevalence counts by subtypes, denoted as $\widehat{N}_{S,it}$, will first be calculated using $\hat{r}_{it}\widehat{N}_{M,it}$, where $\widehat{N}_{M,it}$ is the initial estimate of prevalence count

for total violent crime and $\hat{r}_{it}$ is the estimated ratio[26] of $p_{S,it}$ (the sum of prevalence rates by subtypes) over $p_{M,it}$ (the prevalence rate for total violent crime).

3. **Calculate the national totals:** For each 3-year period, the 3-year average estimates of the national totals of victims for total violent crime (i.e., prevalence counts of total violent crime at the national level) and by violent crime subtypes are calculated based on the NCVS personal-level data that are also used as input data to produce the SAE prevalence estimates for violent crimes.

4. **Adjust the initial prevalence count of total violent crime $\boldsymbol{\hat{N}_{M,it}}$ and the initial estimate of the sum of the prevalence counts by subtypes $\boldsymbol{\hat{N}_{S,it}}$:** $\hat{N}_{M,it}$ from Step 1 are proportionally adjusted to agree with the NCVS 3-year average estimate of total violent crime from Step 2. Similarly, $\hat{N}_{S,it}$ from Step 1 are proportionally adjusted to agree with the sum of the NCVS 3-year average estimates of subtypes from Step 2.

5. **Use raking to adjust the initial prevalence counts by crime subtypes:** Total violent crime can be divided into different sets of subtypes. The initial estimates of each subtype from Step 1 are adjusted using two-dimensional raking so that the estimates agree with the NCVS national estimates of the corresponding subtype from Step 2 (the first margin) and the adjusted $\hat{N}_{S,it}$ from Step 3 (the second margin). These sets include the following:

   - <u>Violent crime by relationship</u>: Crime by strangers, intimate partners, and other non-strangers

   - <u>Violent crime by type</u>: Simple assault and violent crime excluding simple assault (called “serious violent crime” in publications before 2018)

     – Violent crime excluding simple assault: this subtype is furtherly dichotomized into robbery and aggravated assault/rape. Again, the initial estimates of these two subtypes are raked to their national estimates and the estimated sum of prevalence counts of these two subtypes.

6. **Convert the adjusted estimates into prevalence rates:** Adjusted estimates from Steps 3 and 4 are converted to rates per 1,000 based on the population estimates and rounded.

7. **Convert the RMSEs into rates:** The RMSEs are also converted to rates per 1,000 and rounded, based on the results from the model used in estimating each type.

[26] The way in which this ratio is modeled and estimated differs by subtypes. (a) For crime by strangers, intimate partners, and other non-strangers (violent crime by relationships), a univariate model is used for estimating the ratios because of the difficulty of modeling a multivariate model with total violent crime as the other component. (b) For simple assault and violent crime excluding simple assault, ratios are modeled jointly with their denominators (i.e., the prevalence rates of total violent crime) using a multivariate model. (c) For robbery and aggravated assault/rape, ratios are also modeled jointly with their denominators (i.e., the prevalence rates of violent crime excluding simple assault) using a multivariate model.

7.4.2 For Property Crimes

Benchmarking the prevalence rates and numbers for property crimes (implemented through *benchmark_prop_prev_revision.R*) is similar to benchmarking for violent crimes with the following modifications:

1. The prevalence counts for property crimes for each state and 3-year period are estimated by multiplying the small area estimates of the prevalence rates by the 3-year averages of NCVS household estimates instead of the population estimates for in this state and 3-year period.
2. The national totals for total property crime and by subtypes are calculated based on the NCVS house-level data (e.g., *sastoc_prop1416.Rdata*), rather than person-level data. The household-level data are generated based on information described in Section 5.3 and used as input data to produce the small area estimates for property crimes.
3. In Step 4, the total property crime is divided into burglary, motor vehicle theft, and other theft.

7.5 Software Implementation

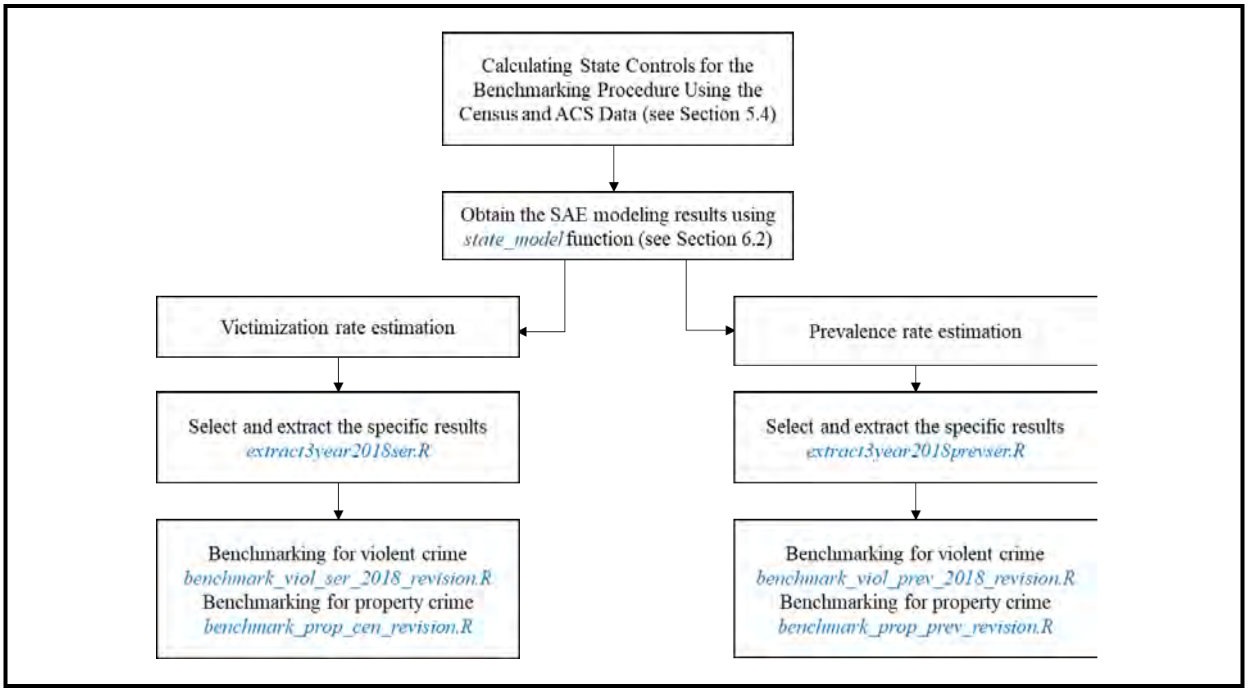
Figure 7.3 displays the process by which the functions are implemented to obtain the small area estimates and perform benchmarking for different types of estimates. This process includes the following four key steps:

1. Obtain the State Controls at both person level and household level based on the census and American Community Survey data and saved as input data files.
2. Get the state-level small area estimates using the *state_model* function. In this step, competing models (e.g., univariate or multivariate models) can be fitted and the results will be saved for consideration.
3. Use *extract3year2018ser_extract.R* (for victimization rate) and *extract3year2018prevser_extract.R* (for prevalence rate) to select and extract the specific results that will be included in the final estimates. These two scripts have a simple structure of
   - loading the files with the results from *state_model*;
   - creating new variables to be recognized by the benchmarking program;
   - extracting the specific contents for the small area estimates and their RMSEs; and, finally,
   - outputting the results in a form that can be used by the benchmarking programs.

4. Apply the benchmarking program to perform benchmarking adjustments for the small area estimates from Step 3. The benchmarking program is different by violent crime and property crime and is different for victimization rates and prevalence rates. The benchmarking R scripts (as provided in the supplemental files) are

    - *benchmark_viol_ser_2018_extract.R*, for victimization of violent crime;
    - *benchmark_viol_prev_ser_2018_extract.R*, for prevalence of violent crime;
    - *benchmark_prop_cen_2018_extract.R*, for victimization of property crime; and
    - *benchmark_prop_prev_cen_2018_extract.R*, for prevalence of property crime.

Figure 7.3 Process to Obtain the SAE Modeling Results and Perform the Final Benchmarking Procedure


Calculating State Controls for the Benchmarking Procedure Using the Census and ACS Data (see Section 5.4)
Obtain the SAE modeling results using *state_model* function (see Section 6.2)
Victimization rate estimation
Prevalence rate estimation
Select and extract the specific results
extract3year2018ser.R
Select and extract the specific results
extract3year2018prevser.R
Benchmarking for violent crime
benchmark_viol_ser_2018_revision.R
Benchmarking for property crime
benchmark_prop_cen_revision.R
Benchmarking for violent crime
benchmark_viol_prev_2018_revision.R
Benchmarking for property crime
benchmark_prop_prev_revision.R

It is worth noting the following issues in the benchmarking R scripts.

- Each script begins by setting pred_start and pred_end for the span of years to be estimated—that is, the lower limit of the first 3-year average and the upper limit of the last. These values should be within the same range as specified in *state_model*.

Note that pred_start should be 1997 or larger, but not equal to 2006,[27] and pred_end should be 2007 or greater.

- The NCVS national totals are calculated based on the same NCVS data files used to fit the SAE models and to get the 3-year small area estimates. The 2004–2006, 2005–2007, and 2006–2008 averages are formed by excluding 2006 results and averaging the other 2 years.

- Each script includes identical functions, including *ncvs_adjust_1(), rake.2(), ncvs_adjust_2(),* and *ncvs_adjust_3()*. These functions generalize the adjustments described in Sections 7.2–7.4. Note that some values from the global environment are referenced in these functions rather than requiring all needed values to be in their argument list.

- Each script outputs three CSV files for each outcome variable (e.g., robbery victimization rate, burglary prevalence rate). These three files contain (a) the rounded rates per 1,000; (b) the rounded RMSEs for the rates per 1,000; and (c) the rounded estimated numbers, respectively.

[27] In modeling the time series, a special approach excluded the NCVS data from 2006 from the state estimates because the implementation of a sample redesign in 2006 seemed to create a large, transient increase in the NCVS estimates for 2006 relative to neighboring years (Rand & Catalano, 2007). The special treatment of year 2006 remains implemented in the NCVS SAE work.

CHAPTER 8. DISCUSSION OF THE STATE-LEVEL NCVS SMALL AREA ESTIMATION RESULTS

8.1 Introduction

At the time of this report, the state-level National Crime Victimization Survey (NCVS) small area estimates of 3-year rolling averages have been produced for crime victimization and prevalence rates for years 2007–2009 to 2016–2018. In this chapter, the results of the NCVS estimation methodology are presented in the following four ways. First, the state-level small area estimates are presented (Section 8.2). Second, the small area estimates are compared to their corresponding direct estimates in the 11 largest states (based on population), which are presented in Moore, Couzens, and Berzofsky (forthcoming; Section 8.3). A summary is provided to illustrate the differences between the small area estimation (SAE) method and the direct estimation method. Third, the 2016–2018 NCVS small area estimates are compared to the estimates based on the 2016–2018 Summary Reporting System (SRS) data collected by FBI's Uniform Crime Reporting (UCR) program for all 50 states and the District of Columbia (Section 8.4). The comparison confirms that the NCVS estimates are higher than the UCR SRS estimates because the NCVS accounts for both reported and unreported crimes, whereas the UCR SRS includes only crimes reported to law enforcement. The results also show that the differences in estimates between NCVS and UCR SRS can vary across states and crime types. This finding suggests that the NCVS small area estimates can be a valuable indicator of crime in addition to the UCR SRS estimates at the subnational level. Finally, the overall SAE procedure and its results are summarized and discussed (Section 8.5).

Summary of Contents in Chapter 8

What is the main point of this chapter?	The state-level National Crime Victimization Survey (NCVS) small area estimates of 3-year rolling averages are presented and compared to their counterparts derived with the NCVS data using the direct estimation approach and to the Uniform Crime Reporting (UCR) Summary Reporting System (SRS) data.
Why is it important?	Comparing the small area estimates with estimates from other sources can help analysts to better understand the reliability and special features of the small area estimates.

What are the key findings in this chapter?	• Only 2.0% of victimization rates and 1.0% of small area estimation (SAE) prevalence rates at the state level have relative standard errors (RSEs) greater than 50%, which is generally interpreted as unreliable. Most of these high RSEs were in states with small populations. • The direct and small area estimates are similar to each other for most years and states among the 11 largest states. • The trend line of NCVS small area estimates is flatter and has less fluctuation than the trend line of the NCVS direct estimates for state-level violent victimization rates in each of the 11 largest states. • As expected, the 3-year state-level NCVS small area estimates are consistently higher than their counterparts derived from the UCR SRS data in 2016–2018. The differences between the NCVS estimates and the UCR SRS estimates also vary across states and by crime types.
What are the implications based on the key findings in this chapter?	• The state-level small area estimates seem to be reliable for relatively large states based on their RSEs and the comparison results between the small area estimates and the direct estimates. • The state-level small area estimates for some small states seem to be unreliable because their RSEs are relatively high (higher than 50%). • One can use both the trend lines of small area estimates and direct estimates to better understand the crime trends for large states. • The NCVS small area estimates can be a valuable indicator of crime in addition to the UCR SRS estimates at the subnational level because the NCVS accounts for both reported and unreported crimes, whereas the UCR SRS includes only crimes reported to law enforcement.

8.2 State-Level NCVS Small Area Estimates for 2007–2018

This report provides explicit details on how to produce the small area estimates of crime victimization and prevalence at the state and substate levels, along with their root mean squared errors (RMSEs). Victimization rates are the total number of times that persons or households were victimized per 1,000 persons or households.[28] Prevalence rates indicate the percentage of persons or households who were crime victims. Both victimization rates and prevalence rates have been calculated for the following 12 crime types:

[28] As discussed previously, only those aged 12 and over living in households or noninstitutionalized group quarters are part of the NCVS population, and this population is the denominator for rates.

- Person-level crime—Total violent crime[29]
    - Violent crime excluding simple assault[30]
    - Assault
        - Simple assault
    - Robbery
    - Intimate partner violence[31]
    - Stranger violence[32]
    - Other relationship violence[33]
- Household-level crime—Total property crime
    - Burglary/trespassing[34]
    - Motor vehicle theft
    - Other theft[35]

The state-level estimates with their RMSEs are provided for victimization rates and prevalence rates in the 50 states and the District of Columbia in the supplemental files, *SAE_Victimization_Rates.xlsx* and *SAE_Prevalence_Rates.xlsx*, respectively.

Each of these two files contains 120 small area estimates for each state. corresponding to ten 3-year rolling averages from 2007 to 2018 and 12 crime types. Because SAE is a model-based method that can have a bias-variance tradeoff (due to model overfitting or underfitting), the small area estimates are more often evaluated in terms of RMSEs rather than in terms of their variances and standard errors (SEs). A relative standard error (RSE) is a measure of precision of the estimate defined as a ratio of the RMSE over the estimate. Some national surveys determine an estimate's reliability using this rule: if the RSE is greater than 50%, the estimate is considered

[29] Excludes homicide because the NCVS is based on interviews with victims.

[30] Includes rape or sexual assault, robbery, and aggravated assault. This category was called "serious violent crime" before 2018.

[31] Includes the subset of domestic-violence victimizations that were committed by intimate partners, which include current or former spouses, boyfriends, or girlfriends.

[32] Includes the subset of violent victimizations that were committed by someone unknown to the victim.

[33] Violent crime where the offender is neither an intimate partner nor a stranger.

[34] This category was called household burglary before 2018. Includes unlawful or forcible entry or attempted entry of places, including a permanent residence, other residence (e.g., a hotel room or vacation residence), or other structure (e.g., a garage or shed), but does not include trespassing on land.

[35] Includes other unlawful taking or attempted unlawful taking of property or cash without personal contact with the victim. Incidents involving theft of property from within the same household would classify as theft if the offender has a legal right to be in the house (such as a maid, delivery person, or guest). If the offender has no legal right to be in the house, the incident would classify as a burglary.

unreliable and should not be reported; if the RSE exceeds 30% but is less than 50%, the estimate can be reported with a flag indicating its precision is questionable; if the RSE is less than 30%, the estimate can be considered reliable.[36] On the basis of the results provided in the supplemental files, only 1.7% of victimization rates and 0.7% of prevalence rates at the state level have RSEs greater than 50%. Most of these high RSEs were in states with small populations.[37] Most RSEs—85.2% of victimization rates and 92.3% of prevalence rates—were less than 30%.

8.3 Comparisons Between NCVS Small Area Estimates and Direct Estimates for 11 Large States From 2007 to 2015

As mentioned in Chapter 1, beginning in 2016 BJS boosted the NCVS core sample in 22 states to produce direct estimates of 3-year rolling averages for these states for certain crime types. To support trend examination of victimization over a longer period of time, Moore et al. (forthcoming) assessed the feasibility of reweighting the pre-2016 NCVS data to produce reliable subnational estimates of violent victimization. They produced state-level direct estimates for the 11 largest states in the country. Estimates[38] were produced for 2007 to 2015 after the NCVS weights were recalibrated to reflect state populations in each of those years. Similarly, Moore et al. also produced direct estimates for the 52 Metropolitan Statistical Areas (MSAs) with at least 1 million persons in 2015. Estimates were produced after the NCVS weights were recalibrated to reflect the MSA populations in each of the years 2007 through 2015. A weight calibration approach based on the generalized exponential model (GEM; Folsom & Singh, 2000) was used in this recalibration process. To increase the sample size and achieve a desirable precision level, data were aggregated over 3 years, and direct estimates of 3-year rolling averages were calculated. The direct estimates were derived for victimization rates of total violent crime as well as its subtypes.

[36] Although this rule for reliability is not standard, it is used by surveys such as the Medical Expenditure Panel Study and the Behavioral Risk Factor Surveillance Survey.

[37] Victimization rates with RSEs greater than 50% were in Arkansas (1), Connecticut (6), Delaware (2), District of Columbia (11), Florida (1), Georgia (5), Idaho (4), Maine (6), Montana (3), New Hampshire (1), New Jersey (18), North Dakota (5), South Carolina (5), South Dakota (6), Vermont (17), Virginia (1), West Virginia (2), and Wyoming (9). Prevalence rates with RSEs greater than 50% were in Idaho (4), Maine (4), Montana (1), New Hampshire (4), North Dakota (3), Vermont (16), and Wyoming (8). The number in parentheses indicates the number of estimates in each state with high RSEs across all crime types and years.

[38] The NCVS data used in this analysis are restricted-use data that can be accessed through U.S. Census Bureau Federal Statistical Research Data Centers (FSRDCs). For more information on FSRDCs, see the following webpage: https://www.census.gov/fsrdc.

For total violent victimization rates, the direct estimates, their associated SEs, the small area estimates, and their associated RMSEs are presented in Table 8.1. The two sets of estimates are compared in Figure 8.1, and their associated SEs and RMSEs are compared in Figure 8.2. The U.S. national total victimization rate is also presented for reference and is calculated directly based on the original NCVS weights before recalibration. Note that, in the table and figures below, a year refers to a 3-year estimate ending in that year. For example, the estimate shown as 2015 represents the estimated rate from 2013–2015.

Table 8.1 Rates of Total Violent Victimization, by State and Estimation Method, 2007–2015

State	Estimate Method	2007–2009		2008–2010		2009–2011		2010–2012		2011–2013		2012–2014		2013–2015	
		Rate	SE[a]	Rate	SE	Rate	SE	Rate	SE	Rate	SE	Rate	SE	Rate	SE
United States, overall	Direct	24.9	0.9	22.3	0.9	21.4	0.9	22.7	0.9	24.0	0.8	23.1	0.9	20.6	0.9
California	Direct	17.3	1.3	17.4	1.9	20.3	2.5	25.8	2.2	26.2	1.8	22.6	1.3	18.4	2.0
	SAE	19.2	2.2	18.3	2.0	19.6	2.0	22.7	2.0	24.0	2.0	22.4	1.9	19.4	1.8
Florida	Direct	18.4	1.4	15.8	1.1	13.0	1.5	11.7	1.6	12.4	1.8	11.1	1.5	11.2	1.8
	SAE	18.3	3.2	15.0	2.9	13.2	2.8	13.6	2.8	14.4	2.7	13.4	2.6	11.4	2.4
Georgia	Direct	18.6	3.7	18.8	2.9	17.5	2.2	16.3	1.8	16.6	2.2	15.8	2.1	11.8	2.7
	SAE	19.0	4.2	16.6	3.8	15.3	3.7	16.1	3.5	17.0	3.4	16.1	3.3	13.4	3.1
Illinois	Direct	32.9	6.2	26.0	5.5	22.6	4.0	21.0	4.9	20.1	4.6	16.1	1.9	14.6	1.7
	SAE	28.8	3.6	24.6	3.3	22.2	3.1	22.2	3.1	22.7	3.0	20.9	2.9	18.1	2.8
Michigan	Direct	26.5	4.6	27.4	7.2	24.6	5.3	22.4	5.3	19.6	2.5	16.3	3.9	16.8	5.0
	SAE	26.2	4.2	23.7	4.0	22.3	3.8	23.0	3.6	23.6	3.5	22.5	3.3	19.9	3.2
New Jersey	Direct	12.4	3.2	9.7	2.3	8.2	2.0	8.0	1.7	9.6	2.7	10.4	2.5	8.8	2.1
	SAE	13.6	3.8	10.7	3.4	9.7	3.3	11.0	3.3	12.7	3.2	12.3	3.1	10.0	3.0
New York	Direct	19.5	2.0	17.5	2.1	18.3	2.4	19.8	1.1	21.0	2.2	22.1	3.4	21.6	5.1
	SAE	19.6	3.0	17.4	2.7	17.1	2.6	18.9	2.6	20.5	2.6	20.5	2.5	18.7	2.3
North Carolina	Direct	25.7	5.7	17.6	2.8	13.5	3.1	12.6	2.7	16.1	3.5	21.8	4.1	17.4	2.6
	SAE	21.9	4.4	18.3	4.1	16.4	3.9	17.1	3.8	18.6	3.6	18.4	3.4	15.8	3.3
Ohio	Direct	30.6	4.0	23.4	4.4	22.3	5.3	20.1	3.4	20.5	2.4	22.5	2.4	21.1	3.1
	SAE	29.9	3.9	25.7	3.6	23.7	3.5	24.4	3.4	25.9	3.3	26.0	3.1	24.4	3.0
Pennsylvania	Direct	29.9	7.8	34.6	9.1	27.3	6.5	29.8	6.6	37.0	5.9	38.7	5.0	32.8	3.9
	SAE	28.5	3.8	26.6	3.6	25.8	3.4	27.9	3.3	30.4	3.2	30.0	3.1	26.8	2.9
Texas	Direct	33.1	3.5	29.1	4.5	24.7	3.0	22.5	3.3	23.6	1.9	21.1	1.6	19.8	1.6
	SAE	29.7	2.8	25.9	2.6	23.2	2.5	22.7	2.5	23.1	2.5	21.7	2.3	19.0	2.2

[a] This table presents standard error (SE) for direct estimates and relative mean squared error (RMSE) for small area estimates.

The direct estimates were obtained based on the standard design-based estimation approach, whereas the small area estimates were obtained based on the model-based SAE approach. These approaches are discussed in Section 2.2. In summary, the main differences between the two estimation methods are as follows:

1. **Sample data used to obtain the estimates:** The direct estimates were based on survey data collected directly from the respondents within each geographic area (i.e., state) and the specified time period (e.g., 20013–2015), whereas the small area estimates were derived from area-level explicit models that use NCVS survey data (as model outcome variables) and UCR SRS data (as model predictors) across different geographic areas from multiple years (e.g., 2007–2018).
2. **Statistical approach used to derive the estimates:** The direct estimates are weighted estimates using the calibrated weights. The Taylor Series Linearization method is used to derive their associated SEs. The small area estimates and their associated RMSEs are based on the dynamic models.
3. **Auxiliary data used to improve the estimates:** The direct estimates incorporated data from only the American Community Survey (ACS) to adjust NCVS national-level weights to reflect the demographic and socioeconomic composition of each area (state); small area estimates incorporated data from the UCR SRS, the ACS, and the decennial census.
4. **Benchmarking procedure used to adjust the estimates:** The direct estimates were not benchmarked; the small area estimates were rigorously benchmarked so that the estimates of crime subtypes agree with the estimate of their aggregated type and the sum of state-level estimates agrees with the published national totals.

As shown in Figure 8.1, despite these differences between the methods used to derive the direct and small area estimates, the direct and small area estimates are similar to each other for most years and states among the 11 largest states (North Carolina and Pennsylvania show more divergence between the estimation methods than other states.) This result provides more confidence for the use of small area estimates, especially in large states. Moreover, because the small area estimates were modeled using a time-series approach with data from multiple years, the trend line of small area estimates tends to be flatter and have less fluctuation than the trend line of direct estimates in each of the 11 states.

Figure 8.2 compares the SEs of the direct estimates with the RMSEs of the small area estimates. In some states, such as Florida, Georgia, and New Jersey, the RMSEs of the small area estimates are consistently higher than the SEs of the direct estimates over time, whereas in Pennsylvania, they are consistently lower than the SEs of the direct estimates over time. In the other states, there is no clear pattern of the differences between the two. The trend line of RMSEs of small area estimates tends to be flatter and have less fluctuation than the trend line of SEs of direct estimates in each of the 11 states.

Because of limitations in producing reliable direct estimates at the state level in the pre-state boost period, the comparison is done for total violent victimization rates only among the 11 large states in this section. However, one can also use the direct estimates from Moore et al. (forthcoming) to evaluate the small area estimates for different violent crime subtypes at state or MSA levels. Although the direct estimates cannot be considered as more reliable and serve as a gold standard when evaluating the small area estimates, they can provide additional insights on whether the small area estimates for a particular state or MSA are reliable. In addition, one can use estimates derived from both the direct and SAE methods to understand crime rates and trends in the large states and MSAs.

Figure 8.1 Rates of Violent Victimization, by State and Estimation Method, 2009–2015

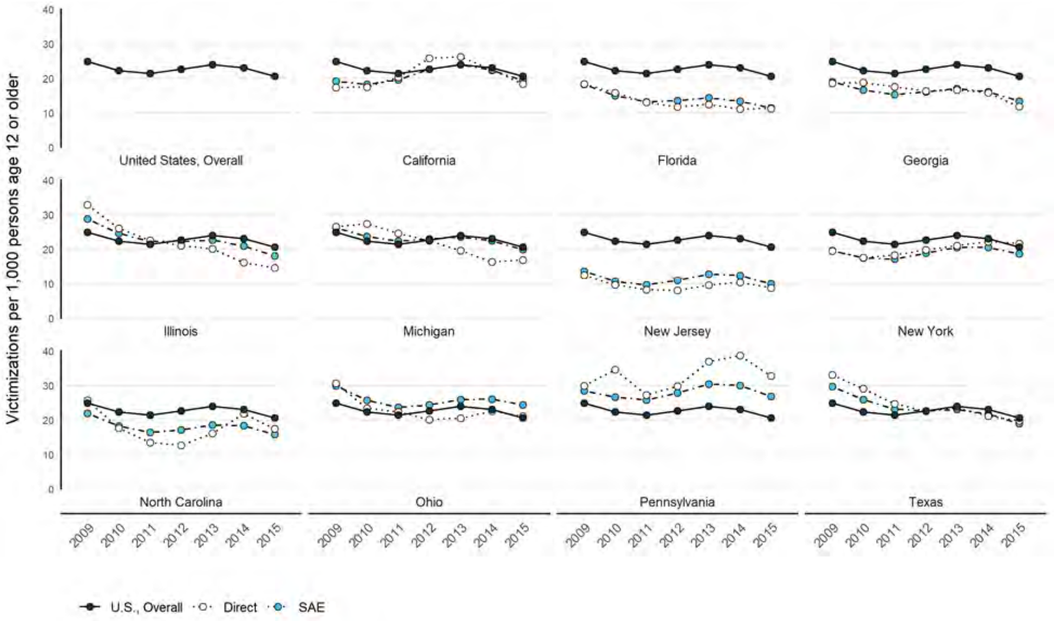
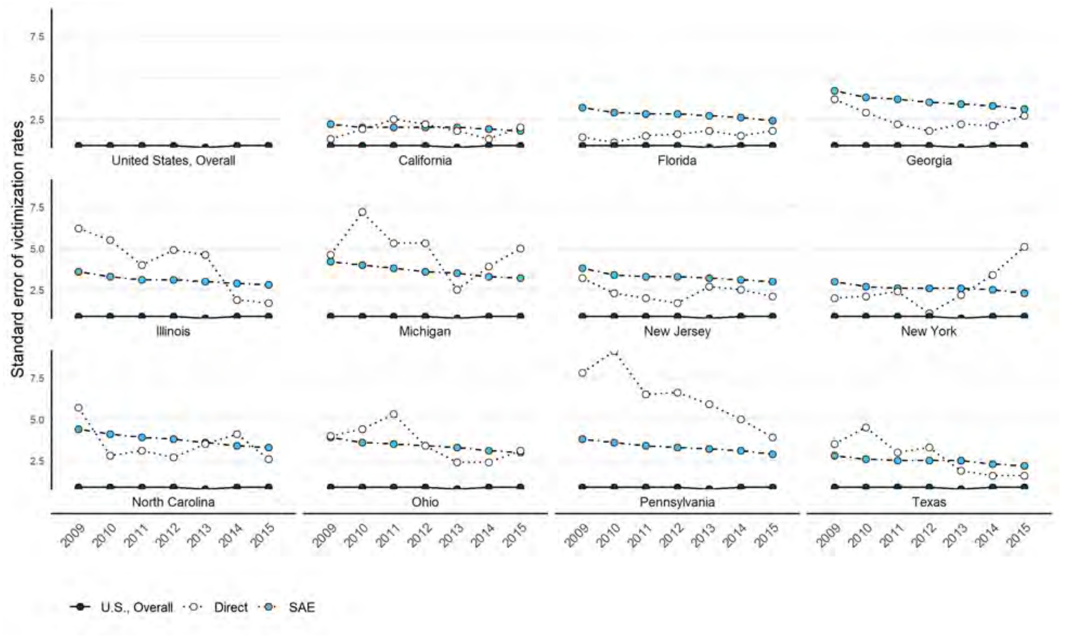


Figure 8.2 SEs and RMSEs of Violent Victimization Rates, by State and Estimation Method, 2009–2015



Note. This plot presents standard error (SE) for direct estimates and root mean squared error (RMSE) for small area estimates.

8.4 Comparisons Between the NCVS Small Area Estimates and the UCR Estimated Crime Data

The NCVS is one of the two main indicators of crime in the United States. Another indicator of crime is the FBI’s UCR, which collect data using two data collection systems, SRS and the National Incident-Based Report System (NIBRS). Law enforcement agencies report crimes through the UCR Program, and state-level estimates are calculated annually based on the UCR SRS data up to the time of the report.[39] The UCR differs from the NCVS in a few key elements:

- The UCR includes only crimes reported to law enforcement
- The UCR includes victims younger than age 12
- The UCR includes victims in institutionalized group quarters

[39] The annual state estimates will be calculated based on the UCR NIBRS data when the SRS sunsets and the coverage of NIBRS data is expanded to the level that enables state-level estimation.

- The UCR estimates only victimizations and not prevalence
- Types of crimes may have different definitions—for example, violent crime in UCR includes homicide but the NCVS does not

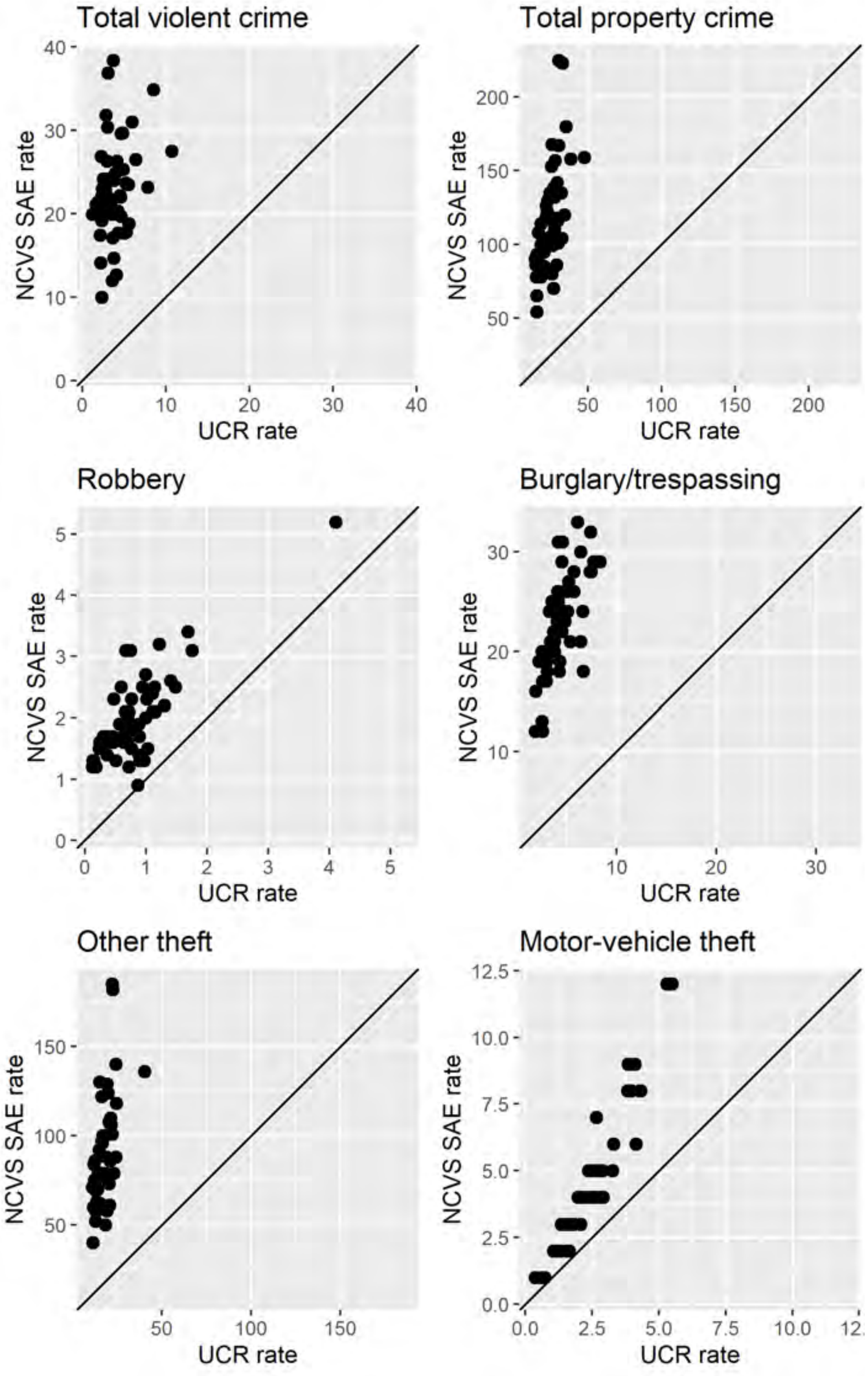
With these differences in mind, the UCR SRS rates are calculated and compared in this section to the NCVS victimization rates for the 3-year period from 2016 to 2018 at the state level. As described in Section 5.2.2, the annual data files of UCR SRS state-level estimates can be downloaded from FBI's Crime Data Explorer website. These data files contain the national- and state-level estimates of UCR SRS data as well as the population size data for each year. With the data files for 2016 through 2018, the UCR SRS 3-year person-level rates are calculated by summing the number of victimizations in the state across the years and dividing by the sum of UCR SRS population for those 3 years. For household-level crimes, an estimate for the number of households is obtained from ACS for each state in each year to calculate the rates. The UCR SRS rates are included in *UCR_estimated_rates.xlsx* in the supplemental files.

For the 50 states and the District of Columbia, their estimated rates in the UCR SRS and the corresponding small area estimates in NCVS are compared in Figure 8.3. Six types of crime are comparable:

- total violent crime
- robbery
- other theft (larceny in UCR SRS)
- total property crime
- burglary/trespassing
- motor vehicle theft

As can be seen in Figure 8.3, the NCVS rates are higher than the UCR SRS rates for all crime types because the NCVS data can reveal the “dark figure” of unreported crime (Skogan, 1977). Robbery and motor vehicle theft rates are most similar between the two sources of crime estimates, which makes sense because they are both highly reported to the police. The variation of the NCVS rates is also higher than the UCR rates across states under all crime types. In addition, the differences between the NCVS rates and the UCR SRS rates vary across different states and crime types. The NCVS rates for states with similar UCR SRS rates can be very different. These findings suggest that the degree of “dark figure” of unreported crime can vary across geographical areas, so the NCVS small area estimates can be a very valuable indicator of crime in addition to the UCR estimates at the subnational level.

Figure 8.3 Comparison of UCR and NCVS SAE 3-Year Victimization Rates, 2016–2018



8.5 Summary

The SAE method has been recognized as a useful statistical tool to inform policy decisions in the absence of sufficient direct sample data. Although the NCVS core sample has been boosted in 22 large states to obtain key direct estimates in these states, the NCVS sample data do not guarantee enough sample to facilitate direct estimation in other geographical areas or for certain crime types. Therefore, an SAE procedure has been developed to fill these gaps and produce estimates for all states and some large counties and MSAs.

A series of R functions, data files, and reports (including this report and Fay, 2021) have been developed to help analysts understand and implement this procedure on their own. Given the complexity of this procedure, one should gain a solid understanding of the basic components under this SAE procedure based on these materials, including

- the NCVS sample design and how to accommodate it in the variance-covariance matrix,
- the auxiliary information from the UCR SRS data and the census and ACS data, and
- the SAE techniques of using dynamic models and benchmarking adjustments to produce the final estimates.

On the basis of the assessment results described in this chapter, the small area estimates derived from this SAE procedure seem to be reliable for relatively large states. Their values are close to the direct estimates generated for the 11 largest states and their RSEs are relatively low. In addition, the comparison between the NCVS small area estimates and the UCR estimates reveals that the crime estimates between the two data sources are different, and their difference can vary across states and by crime types. This result indicates the importance of using NCVS small area estimates to understand crimes at subnational levels that cannot be discovered using the UCR data. The SAE procedure is expected to be used for NCVS in the future to provide informative crime estimates at subnational levels. It is also hoped that the R functions and innovations developed for this procedure can contribute to the research and implementation of SAE methods for other surveys.

CHAPTER 9. REFERENCES

Biderman, A. D., Cantor, D., Lynch, J. P., & Martin, E. (1986). *Final report of research and development for the redesign of the National Crime Survey* (prepared for the Bureau of Justice Statistics). Washington, DC: Bureau of Social Science Research.

Blumberg, S. J., Luke, J. V., Ganesh, N., Davern, M. E., & Boudreaux, M. H. (2012). Wireless substitution: State-level estimates from the National Health Interview Survey, 2010–2011. *National Health Statistics Reports, 61*, 1–16. https://pubmed.ncbi.nlm.nih.gov/24988815/

Bureau of Justice Statistics. (2017, December 8). *National Crime Victimization Survey, 2016: Technical documentation* (NCJ 251442). https://www.bjs.gov/content/pub/pdf/ncvstd16.pdf

Fay, R. E. (2021). *Constructing and Disseminating Small Area Estimates from the National Crime Victimization Survey, 2007–2018.*

Fay, R. E., & Diallo, M. S. (2012). Small area estimation alternatives for the National Crime Victimization Survey. In *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section, American Statistical Association* (pp. 3742–3756). http://www.asasrms.org/Proceedings/index.html

Fay, R. E., & Diallo, M. S. (2015a). *Developmental estimates of subnational crime rates based on the National Crime Victimization Survey* (NCJ 249238). Washington, DC: Bureau of Justice Statistics. https://www.ncjrs.gov/app/publications/abstract.aspx?id=271379

Fay, R. E., & Diallo, M. S. (2015b). *sae2: Small area estimation: Time-series models, R package version 0.1-1*. URL: https://CRAN.R-project.org/package=sae2

Fay, R. E., & 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*, 269–277. https://doi.org/10.2307/2286322

Fay, R. E., & Li, J. (2011). Predicting violent crime rates for the 2010 Redesign of the National Crime Victimization Survey (NCVS). In *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section, American Statistical Association* (pp. 1663–1676). http://www.asasrms.org/Proceedings/index.html

Fay, R. E. , & Li, J. (2012, January 10–12). *Rethinking the NCVS: Subnational goals through direct estimation*. Federal Committee on Statistical Methodology Research Conference, Washington, DC, USA. https://nces.ed.gov/FCSM/2012research.asp#TuesdayAM

Fay, R. E., Planty, M., & Diallo, M. S. (2013). Small area estimates from the National Crime Victimization Survey. In *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section, American Statistical Association* (pp. 1544–1557). http://www.asasrms.org/Proceedings/index.html

Federal Bureau of Investigation. (2020, May 7). *Criminal Justice Information Services Division Uniform Crime Reporting Program 2019.2 National Incident-Based Reporting System user manual*. Washington, DC: U.S. Department of Justice. https://www.fbi.gov/services/cjis/ucr/data-documentation

Folsom, R. E., Jr., & Singh, A. C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. In *JSM Proceedings, Survery Research Methods Section* (pp. 598–603). Alexandria, VA: American Statistical Association. http://www.asasrms.org/Proceedings/papers/2000_099.pdf

Ghosh, M., & Rao, J. N. K. (1994). Small area estimation: An appraisal. *Statistical Science, 9,* 55–76. https://doi.org/10.1214/ss/1177010647

Groves, R. M., & Cork, D. L. (2008). *Surveying victims: Options for conducting the National Crime Victimization Survey* (Panel to Review the Programs of the Bureau of Justice Statistics, National Research Council, Committee on National Statistics and Committee on Law and Justice, Division of Behavioral and Social Sciences and Education). Washington, DC: National Academies Press. https://www.nap.edu/catalog/12090/surveying-victims-options-for-conducting-the-national-crime-victimization-survey

Kong, A. Y., & Zhang, X. (2020). The use of small area estimates in place-based health research. *American Journal of Public Health, 110,* 829–832. https://doi.org/10.2105/ajph.2020.305611

Li, J., Diallo, M. S., & Fay, R. E. (2012, January 10–12). *Rethinking the NCVS: Small area approaches to estimating crime*. Presented at the 2012 Federal Committee on Statistical Methodology Conference, Washington, DC. https://nces.ed.gov/FCSM/2012research.asp

Liu, B., Parsons, V., Feuer, E. J., Pan, Q., Town, M., Raghunathan, T. E., Schenker, N., & Xie, D. (2019). Small area estimation of cancer risk factors and screening behaviors in US counties by combining two large national health surveys. *Preventing Chronic Disease*, *16*, Article E119. https://doi.org/10.5888/pcd16.190013

Lumley, T. (2020). *Survey: Analysis of complex survey samples*. R package version 4.0. https://cran.r-project.org/web/packages/survey/survey.pdf

Luzi, O., Solari, F., & Rocci, F. (2018). A study of small area estimation for Italian structural business statistics. *Journal of Official Statistics, 34,* 543–555. https://doi.org/10.2478/jos-2018-0025

Lynch, J. P., & Addington, L. A. (Eds.). (2006). *Understanding crime statistics: Revisiting the divergence of the NCVS and the UCR* (Cambridge Studies in Criminology). Cambridge, United Kingdom: Cambridge University Press. https://doi.org/10.1017/cbo9780511618543

Moore, A., Couzens, G. L., & Berzofsky, M. (forthcoming). *Assessment of crime victimization in large states and MSAs through direct estimation with recalibrated weights: Evaluation and methodology*. Bureau of Justice Statistics.

Morgan, R. E., & Kena, G. (2018, October). *Criminal victimization, 2016: Revised* (NCJ 252121). Bureau of Justice Statistics. http://www.bjs.gov/content/pub/pdf/cv16re.pdf

Penick, B. K. E., & Owens, M. E. B. (1976). *Surveying crime* (Panel for the Evaluation of Crime Surveys, Committee on National Statistics, Academy of Mathematical and Physical Sciences). Washington, DC: National Academy of Sciences.

Rand, M., & Catalano, S. (2007, December). *Criminal victimization, 2006* (NCJ 219413). Bureau of Justice Statistics. http://www.bjs.gov/content/pub/pdf/cv17.pdf

Rao, J. N. K. (2003). *Small area estimation*. New York, NY: John Wiley & Sons.

Rao, J. N. K., & Molina, I. (2015). Empirical best linear unbiased prediction (EBLUP): Theory. In *Small area estimation* (pp. 97–122). John Wiley & Sons. https://doi.org/10.1002/9781118735855

Rao, J. N. K., & Yu, M. (1992). Small area estimation by combining time series and cross-sectional data. In *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section, American Statistical Association* (pp. 1–9). http://www.asasrms.org/Proceedings/index.html

Rao, J. N. K., & Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics*, *22*, 511–528. https://doi.org/10.2307/3315407

Shook-Sa, B. E., Lee, P., & Berzofsky, M. (2015). *Assessing the coverage and reliability of subnational geographic identifiers in the NCVS public-use file* (NCJ 249467). Washington, DC: Bureau of Justice Statistics. https://www.ncjrs.gov/pdffiles1/bjs/grants/249467.pdf

Skogan, W. G. (1977). Dimensions of the dark figure of unreported crime. *Crime & Delinquency*, *23*, 41–50. https://doi.org/10.1177/001112877702300104

Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T., & Rojas-Perilla, N. (2018). From start to finish: A framework for the production of small area official statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*, 927–979. https://doi.org/10.1111/rssa.12364

U.S. Department of Justice. (2004). *The nation's two crime measures* (NCJ 122705). Washington, DC. https://www.bjs.gov/index.cfm?ty=pbdetail&iid=802

Williams, R., Heller, D., Couzens, G. L, Shook-Sa, B., Berzofsky, M., Smiley-McDonald, H., & Krebs, C. (2015). *Evaluation of direct variance estimation, estimate reliability, and confidence intervals for the National Crime Victimization Survey* (NCJ 249242). Washington, DC: Bureau of Justice Statistics. https://www.bjs.gov/content/pub/pdf/edveercincvs.pdf