This report was prepared by Westat, Inc., using federal funding provided by the Bureau of Justice Statistics.

Document Title: Constructing and Disseminating Small Area Estimates from the National Crime Victimization Survey, 2007–2018

Authors: Robert E. Fay, Westat, Inc.

Document No.: NCJ 300603

Publication Date: April 2021

Award No.: This project was supported by award number 2017-BJ-CX-K030.

Abstract:

This report was produced by Westat, Inc., for BJS under award number 2017-BJ-CX-K030. It builds upon a previously released report that combined data from the National Crime Victimization Survey (NCVS) with small area estimation methods and data from the FBI's Uniform Crime Reports to produce estimates for states for 1999–2013 and for some large counties and substate areas for 1998–2012. This report features state-level estimates for an expanded set of characteristics that have been produced as 3-year averages for 2007–2018. The report includes displays of the state estimates accompanied by a more simplified description for the interested public. The majority of the report focuses on the software and its underlying methods.

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 report refers to software developed by the author to carry out the analysis described for NCVS data. Persons interested in accessing the software to analyze NCVS data for their own research purposes should contact the U.S. Bureau of the Census regarding their Federal Statistical Research Data Center program at https://www.census.gov/about/adrm/fsrdc.html. As of the release of this report, access to NCVS data is limited to years prior to 2016. Persons interested in copies of the software for adapting to their own small area estimation projects not involving NCVS data should send a request to the Bureau of Justice Statistics at askbjs@usdoj.gov.

This page intentionally left blank.

Constructing and Disseminating Small Area Estimates from the National Crime Victimization Survey, 2007-2018

Robert E. Fay, Westat, Inc.

Final report for award #2017-BJ-CX-K030
April 2021

Contents

Executive Summary

Under a previous cooperative agreement between the Bureau of Justice Statistics and Westat, the National Crime Victimization Survey was combined with small area estimation methods and data from the FBI's Uniform Crime Reports to produce estimates for states for 1999-2013 and for some large counties and substate areas for 1998-2012.

This project had two major goals: to systematize and document the software to the degree that other researchers could extend its use into the future and to illustrate its use by producing new estimates. A large part of the research was conducted under the confidentiality restrictions of the Census Bureau's Title 13, and some of the products of the research remain at the Census Bureau for future use by others permitted access to the NCVS internal files. By agreement, these have been shared with the researchers at the Research Triangle Institute collaborating with BJS, who also have access to the NCVS confidential data.

State-level estimates for an expanded set of characteristics have been produced as 3-year averages for 2007-2018 and approved for release. The report includes displays of the state estimates that may be used in a public release of the results accompanied by a more simplified description for the interested public. The majority of the report focuses on the software and its underlying methods.

Introduction and Background

The National Crime Victimization Survey (NCVS) is a household survey conducted to measure crime on an annual basis as reported by its victims. Because a substantial fraction of the occurrence of crime is never reported to the police, the survey provides a more complete account of the national impact of crime than the FBI's Uniform Crime Reports (UCR). An annual Bureau of Justice Statistics (BJS) publication, *Criminal Victimization*, summarizes key findings and documents the classification of crimes used by the survey. The survey also supports specialized studies and provides public use files for research purposes. To protect the confidentiality of respondents, the public use files contain only limited geographic detail, such as indicating the four census regions (Northeast, Midwest, South, and West), but not, for example, state identifiers.[1]

The geographic detail published from the survey has also been limited. An exception is the report *Developmental Estimates of Subnational Crime Rates based on the National Crime Victimization Survey* (Fay and Diallo, 2015a), which was the outgrowth of a previous cooperative agreement between BJS and Westat (2008-BJ-CX-K067). The research implemented small area models by relating the NCVS estimated crime rates over time to UCR statistics. The models were developed using geographic identifiers on the Census Bureau's internal NCVS files, and the results cannot be reproduced from publicly available files. The models provided State-level estimates of incidence rates in the form of thirteen overlapping 3-year intervals from 1999-2001 to 2011-2013.

Similarly, estimates for large counties and CBSAs were provided for thirteen 3-year intervals 1998-2000 to 2010-2012, https://www.bjs.gov/index.cfm?ty=tp&tid=911. The estimates included the average annual incidence of violent crime in each 3-year interval, both for total violent crime and disaggregated into simple assault, robbery, and a remaining category of combined assault, which groups together aggravated assault, rape, and sexual assault. Violent crime was also separately disaggregated into crimes by strangers, crimes by intimate partners, and crimes by others. Total property crime was disaggregated into burglary, motor vehicle theft, and larceny, that is, all other theft. These characteristics were selected for small area modeling in consultation with BJS. The corresponding national estimates have been featured prominently in the series *Criminal Victimization.*

As is typical of many small area projects in their initial phases, the software developed under the first cooperative agreement evolved. Some components were more general than others; the most general and mature components of the software implementing the statistical models were released publicly as the sae2 package written in the R language. The package remains available

[1] The entire series of *Criminal Victimization* may be found on the BJS website, https://www.bjs.gov/index.cfm?ty=pbse&sid=6. The NCVS public use files are available through the National Archive of Criminal Justice Data (NACJD) https://www.icpsr.umich.edu/icpsrweb/content/NACJD/index.html within the Inter-collegiate Consortium for Political and Social Sciences (ICPSR) of the University of Michigan.

from cran.r-project.org to researchers with small area applications involving a time series of survey estimates. But most other aspects of the software were specific to the application to NCVS. For example, the software reflected the NCVS sample design at the time and did not anticipate future design changes.

The Census Bureau redesigns the NCVS sample every ten years based on results from the most recent census, but the 2016 redesign based on the 2010 Census substantially altered the existing design with the goal of producing publishable estimates for the 22 most populous states. Although the small area software produced under the previous cooperative agreement was easily modified to cover the NCVS design up to 2015, the software could not accommodate the features of the 2010 redesign implemented in 2016 without requiring considerable modification. (See Morgan and Kena, 2017, for a high-level summary of the design changes.)

To facilitate the production of estimates for additional years and for new characteristics, BJS funded a new cooperative agreement (2017-BJ-CX-K030) in 2017 to extend and document the small area methods for NCVS. This report covers this work and reflects two primary goals. The first is to update and expand the scope of previously published estimates. The report presents state-level incidence estimates for the same types of crimes as before but adds prevalence rates for each type. Prevalence rates estimate the proportion of persons victimized by the type of crime during the year, regardless of the number of times. NCVS prevalence estimates are comparatively new and were based on research by Lauritsen, et al. (2012) and first reported in the *Criminal Victimization* report for 2013 (Truman and Langton, 2014). The new state-level estimates of incidence and of prevalence are for ten 3-year intervals from 2007-2009 to 2016-2018.

The second goal is to describe the software now supporting the new estimates in order to guide other researchers to extend small area estimates in future years and to model additional characteristics. Some sections of the report describe the design and characteristics of the software. Other sections provide further operational detail about the software for those who wish to use the software in their own research.

Under the previous cooperative agreement, estimates for the period 1998-2012 were also released for the 65 largest counties and for the 51 largest metropolitan areas as defined by the 2010 census. The software has been generalized to support updating estimates for these substate areas. As of yet, however, the Inter-collegiate Consortium for Political and Social Research (ICPSR) has not published the 2015 county-level UCR estimates used as input to the models, even though the ICPSR has released comparable 2016 results. This report suggests waiting until both the 2015 and 2017 county-level estimates are available from ICPSR in order to produce small area estimates for 2007-2017. A later section of the report describes the statistical considerations in modeling substate areas and the software provided for this purpose.

The December 2015 BJS publication was an outgrowth of one of two reports (Fay and Diallo, 2015b) submitted to BJS at the end of the final continuation of the first cooperative agreement. BJS staff assisted in the revision of the report for the December release. A second report (Fay and Diallo, 2015c) provided more detail on technical aspects of the previous research.

The origins of the previous small area work can be traced to two of the recommendations in a report by the National Academy of Sciences (NAS) of the National Research Council (Cork and Groves, 2008). One was to investigate possible modifications to the sample design to improve the efficiency of the survey estimation, and the other was to increase the utility of the NCVS by expanding its geographic scope through modeling. The previous cooperative agreement with Westat (2008-BJ-CX-K067) was initially focused on the first of the two recommendations. Another agreement with Westat, begun at approximately the same time, reviewed options to produce subnational estimates, primarily for states and metropolitan areas, through either (1) expansion or reallocation of the NCVS sample, (2) indirect estimation based on modeling, or (3) supplemental, lower-cost survey collections. Recommendations from this effort were reported by Cantor et al. (2010).

Under the previous agreement (2008-BJ-CX-K067), an investigation of the seemingly high sampling variance for the estimates of the national incidence of crime uncovered large contributions from two sources:

1. the sampling of primary sampling units (PSUs) in non-certainty strata, and
2. reports of multiple incidents by a relatively small proportion of victims.

These two findings largely accounted for why the estimated sampling errors of NCVS incident rates seemed considerably higher than the simple random sampling variance of a proportion with the same sample size.

In addition, research on year-to-year correlations formed a basis for estimating the effect of averaging subnational estimates over time. Cantor et al. (2010) pointed out that, in general, forming averages over time improves the stability of the estimates compared to their individual annual values. But the amount of improvement depends on the correlations between years, with low correlations leading to more improvement compared to higher correlations.

Under the previous agreement (2008-BJ-CX-K067), analysis of NCVS internal files found that year-to-year correlations were considerably higher in non-certainty PSUs than in the certainty PSUs. This finding was consistent with the finding above of the large contribution of between-PSU variance from non-certainty PSUs. The findings provided an initial indication of how much an increase in the NCVS sample in selected states could improve the average state estimates when averaged over a given number of years. A model was developed to predict how supplementing the NCVS sample could improve the precision of selected state estimates, including the additional boost in precision from transitioning some PSUs from non-certainty to certainty status in the design (Fay and Li, 2012).

During this phase of the previous agreement, exploratory research also investigated subnational correlations between NCVS estimates of crime and statistics from the UCR, both at the level of large, certainty counties and at the state level. The goal was to inform strategies for stratifying the sample for the 2010 redesign. The UCR estimates of rape and sexual assault were more predictive of both simple and aggravated assault in NCVS than UCR statistics on aggravated assault. UCR robbery was correlated with NCVS estimates of robbery. There was a correspondence between NCVS and UCR estimates of property crime: UCR motor vehicle theft, UCR burglary, and UCR larceny each predicted the analogous characteristics in the NCVS. The UCR reporting of murder did not show a strong relationship to NCVS violent crime. Similarly, tenure, whether the household owns or rents the house, did not show a correlation at the area level, in spite of a strong relationship at the household level, with renters reporting higher rates of violent crime in the NCVS (Fay and Diallo, 2015c).

Two continuations of the previous cooperative agreement supported the examination of small area estimation methods, as suggested by the NAS and then studied by Cantor et al. (2010). Among other alternatives, Cantor and his colleagues suggested small area estimation based on the time series model proposed by Rao and Yu (1994). Under the previous agreement (2008-BJ-CX-K067), this model and a modification of it termed the *dynamic model*, (Fay and Diallo, 2012) showed promise in combining information across time to improve the prediction, beyond a simple cross-sectional model for a fixed year. The two models and their possible application were elaborated in a series of papers, including a multivariate generalization (Fay, Planty, and Diallo, 2013) enabling the modeling of related types of crime.

The final report from the previous agreement described two issues to be treated in more detail in this report, estimating sampling variances and covariances for the direct estimates based on the NCVS sample, and benchmarking the state small area estimates to be logically consistent and to agree with the national NCVS estimates. In modeling the time series, a special approach excluded the NCVS data from 2006 from the state estimates (Fay and Diallo, 2015b, 2015c), 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 and Catalano, 2007). The special treatment of year 2006 remains implemented in the current software. A related issue will be discussed in detail in this report, where some of the 2016 NCVS data as originally published exhibited similar transient spikes (Morgan and Kena, 2017). In consultation with BJS, the Census Bureau revised the 2016 estimates, excluding some but not all of the 2016 data (Morgan and Kena, 2018). The revised 2016 estimates were used in the current models.

As additional background, the next section describes features of the 2010 NCVS sample redesign influencing the approach to the state-level modeling. Section 4 introduces the state-level estimates for 2007-2018, which are shown graphically in Appendices A and B. Section 5 presents an overview of the software system that produced the state estimates. Subsequent sections describe aspects of the software system in more detail: section 6 describes the preparation of UCR data as input; section 7, covariance modeling of the direct NCVS state estimates; section 8 how models are specified, including the models used for the 2007-2018 estimates; and section 9, the benchmarking of the estimates as a final step.

The 2010 NCVS Sample Redesign

Previous sample redesigns of the NCVS had the primary purpose of improving the efficiency of the survey's national estimates by incorporating information from the most recent U.S. decennial census. In addition to this goal, the redesign based on the 2010 census added the objective of producing estimates for 3-year averages of the incidence of violent crime for 22 individual states. The new design consequently required an expanded sample size, beginning in 2016. The 22 supplemented states were those with the largest populations: 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.

To produce state-level estimates for the 22 supplemented states, the first-stage stratification of the primary sampling units (PSUs) was state-based, that is, sampling strata did not cross state boundaries for these states. In the NCVS design, each PSU comprises one or more counties. Typically, the largest metropolitan areas in each supplemented state became self-representing strata, included with certainty. In most of the supplemented states, one or more non-certainty strata were also defined, each with two or more non-self-representing PSUs. A single PSU was randomly sampled from each of the non-certainty strata. In the redesign, the four largest states, California, Florida, New York, and Texas, each received roughly the same sample size as they would have been entitled to in a national design. In the remaining supplemented states, the supplementation significantly increased the sample size, with an associated decrease in the average survey weight.

Although the redesign did not attempt to support reliable estimation for the remaining states and the District of Columbia, the Census Bureau decided to extend state stratification to them, guaranteeing some NCVS sample in every state. This design decision ensured flexibility to expand the sample at a future date to supplement additional states beyond the first 22 without disrupting the sample in other states. In some cases, a state might be represented by a single PSU sampled from a stratum of non-self-representing PSUs covering the entire state. In the non-supplemented states, the allocated sample size was determined by national estimation goals. Consequently, the survey weights in these states were approximately equal across this group of states and roughly similar to those of California, Florida, New York, and Texas.

The transition from the PSU design based on the 2000 census to the design based on the 2010 census occurred in January 2016. In many cases, PSUs from the earlier design remained in the 2010 redesign, although their selection probabilities may have changed. For example, a non-self-representing PSU in a supplemented state might have become self-representing in the redesign, particularly in the 22 supplemented states. The first-stage samples of PSUs can be regarded as statistically independent between the 2000 and 2010 designs. Sampled PSUs during the period 2015-2017 can be classified as outgoing if they are in the 2000 design only, incoming if in the 2010 design only, or continuing if in both designs.

The transition between housing units sampled from the 2000-based frame and those based on the updated frame based on the 2010 census extended over three years. In 2015, the first 2010-based samples were introduced in continuing PSUs. At the beginning of 2016, all sampled housing units in outgoing PSUs were dropped from the design, while a sample drawn from the 2010 frame was introduced for incoming PSUs. By the end of 2017, 2000-based sample in continuing areas completed their final interviews, so that the sample in 2018 was drawn entirely from the 2010 frame (BJS, 2017).

State Estimates for 2007-2018

The previous state estimates for 1999-2013 covered the incidence of violent crime by type, of violent crime by relationship to the perpetrator, and of property crime by type. During the initial research, the models were also tested on the period 1998-2012. The suggestion at the time was that future NCVS small area estimates might be offered for sliding 15-year windows, reflecting a tradeoff between the currency of the data and having sufficient data to estimate the coefficients of the models effectively.

The new set of estimates include the same set of crimes as before, but for the 12-year span 2007-2018. A possible path forward for future estimates would be to maintain a sliding 12-year window. For example, the next 12-year window will be 2008-2019. The motivation for the change to a 12-year window was to eliminate further involvement of the problematic year 2006. The additional data from the sample expansion will approximately offset the variability in estimating the coefficients of the model from the shortened span.

The plots in Appendix A show the estimates by year from both the previous 1999-2013 set and the current one, with overlaps for five 3-year periods, 2007-2009, 2008-2010, 2009-2011, 2010-2012, and 2011-2013. The incidence is stated as rates per 1,000. The plot for violent crime by type in Alabama is typical (***Figure 1***). Here, slight discrepancies can be seen in the predictions for total violent crime and simple assault during the overlap. Most of the states display relatively minor differences during the overlap between the estimates based on the 1999-2013 and the 2007-2018 data.

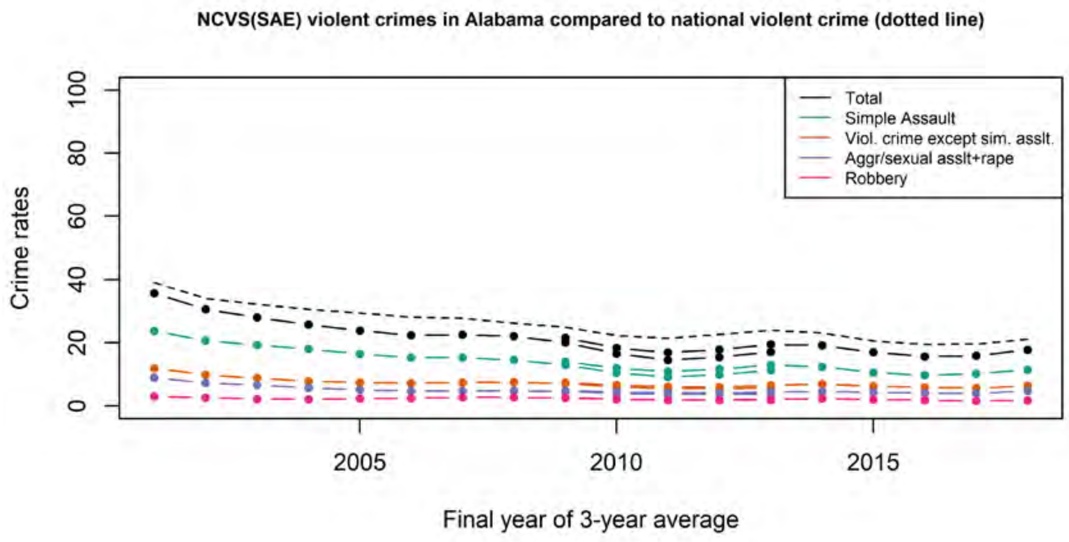

Figure 1. Small area estimates of violent crime by type for 1999-2003 and 2007-2018 in Alabama from Appendix A.

As a reference, the graphs in Appendix A include the 3-year average of the national rates for the violent crime rate or the property crime rate as a dotted line. In ***Figure 1***, Alabama's estimates for violent crime during the entire period fall somewhat below the national average during the period covered by the graph.

In a few states, including Hawaii and Nebraska, there is more of a disconnect between the two sets of estimates. The differences in the District of Columbia appear to be the most sizeable. The largest discrepancies are restricted to small states, which were not included among the 22 supplemented states. Even in most small states, the estimates align well, as illustrated by Alabama, but in small states the estimates depend primarily on the estimated coefficients for the UCR, and in states with atypical UCR values the effect on the predictions is notable.

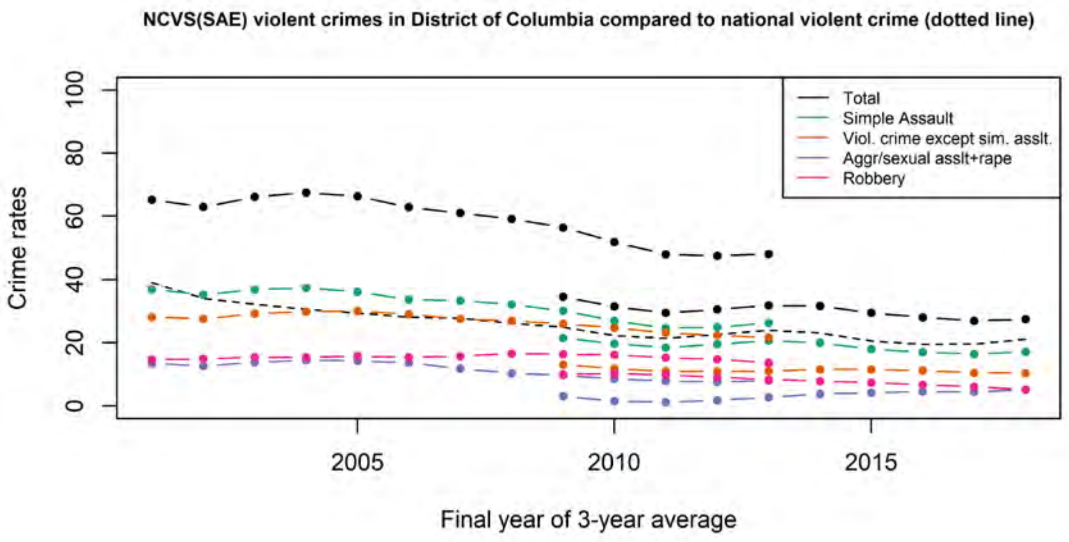

Figure 2. Small area estimates of violent crime by type for 1999-2003 and 2007-2018 in the District of Columbia, from Appendix A.

Because small area estimates trade-off bias and variance, they are more often evaluated in terms of their root mean square errors (RMSE) rather than only their variances. Root mean square errors can be compared to the standard errors for direct survey estimates, so that a relative root mean square error of 2 per 1,000 is competitive with a standard error of 2 per 1,000 for a direct estimate. Root mean square error estimates accompany the small area estimates in the spreadsheets.

Figure 3 summarizes the RMSE estimates by averaging them for three groups of states: the 4 largest states namely California, Florida, New York, and Texas, which had substantial sample sizes during the entire period, the remaining 18 largest states benefitting from the supplementation beginning in 2016, and the remaining smaller states. The figure suggests that the supplementation benefitted the small area estimates for even the smallest states, as the increased sample sizes may have improved the precision of the regression coefficients in the model. Nonetheless, the small states remain less precisely estimated than the supplemented states by the end of the period.

Prevalence estimates are presented as percentages in Appendix B for the period 2007-2018. ***Figure 4*** presents a parallel analysis of RMSE estimates for the prevalence. Here, the largest gains appear to the group of supplemented states except for the largest four.

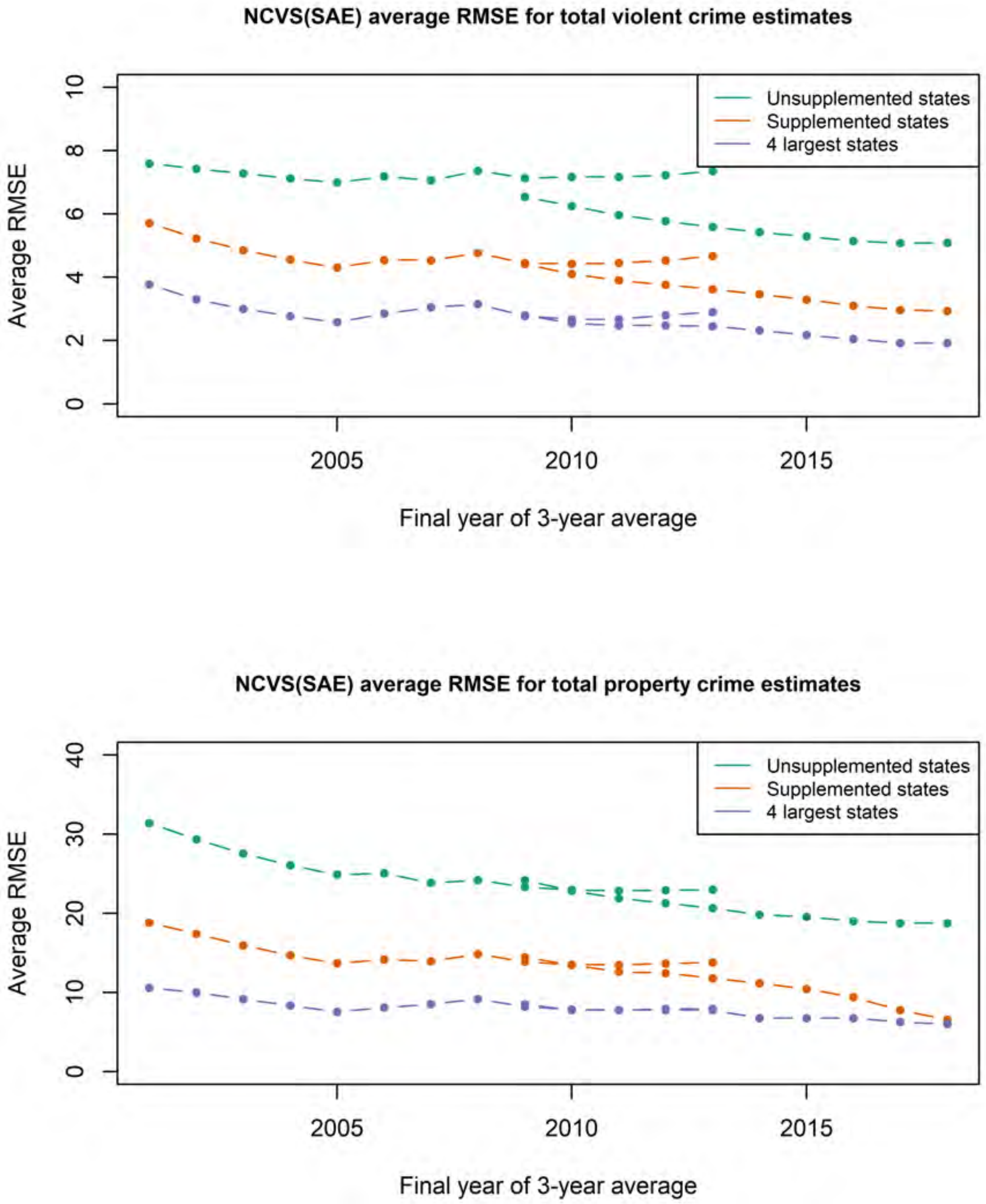

Figure 3. Average RMSE estimates for the state-level small area incidence estimates.

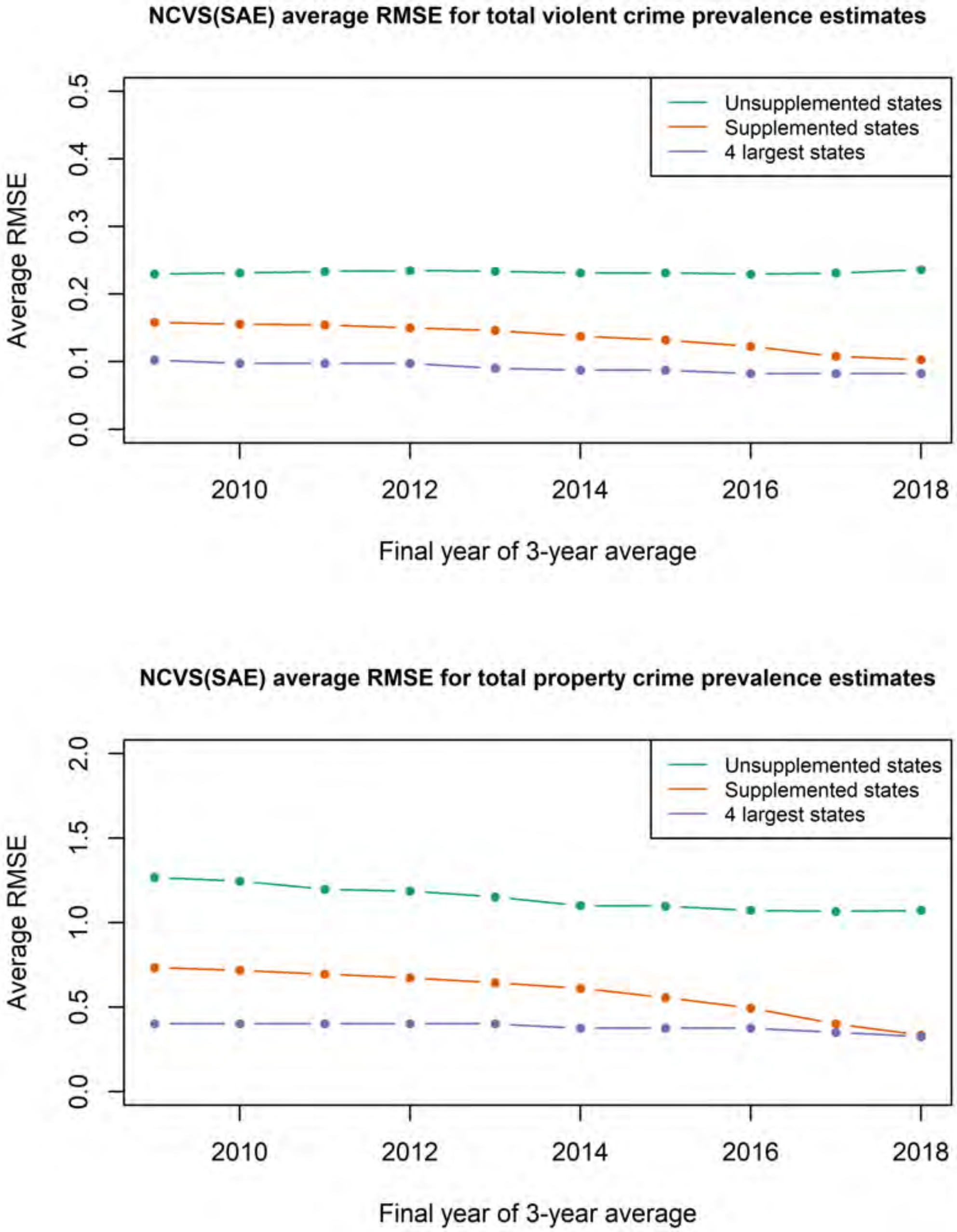

Figure 4. Average RMSE estimates for the state-level small area prevalence estimates.

Design of the NCVS Software for State-Level Small Area Estimation

Overview

A set of generalizable programs produced the accompanying state small area estimates for 2007-2018. Most components of the system were also tested on the periods 2001-2015 and 2003-2017, although these interim results were not submitted to the Census Bureau's Disclosure Review Board for release. The software is designed to be easily run in the future to incorporate additional years of NCVS data under the 2010 design. The software will also permit researchers to investigate small area estimates for additional NCVS characteristics.

The software is primarily written in R, with a set of SAS programs to interface with the Census Bureau's SAS-based internal files. Communication between SAS and R is implemented by writing data out as character .csv files, which R can readily read. This approach has proven satisfactory.

The software is designed to extract necessary information from the Census Bureau's internal files, implement the small area models, and produce the results in the form that was submitted to the Disclosure Review Board for release. The process of small area estimation can be described as a series of steps:

1. A set of 14 SAS programs extracts data from the Census Bureau's internal SAS data files for all years from 1998 to 2018. The outputs are .csv files at the household, person, and incident levels for each year.
2. A SAS program extracts information from three internal SAS data sets (one each at the household, person, and incident level) used to form the 2016 bridge data. The files include identifiers and new weights to indicate how to combine the 2015 and original 2016 data sets into the 2016 bridge file used for the revised estimates. The outputs are .csv files.
3. A short R program creates R data frames from the .csv files from steps 1 and 2.
4. A short R program combines the 2015 R data frames, the original 2016 R data frames, and the files from steps 2 and 3 into an R data frame for the 2016 bridge file with the same variables as the other years.
5. Three R functions translate internal SAS names into the names used on the public use versions of the files available from ICPSR at the University of Michigan. To add new variables, the functions must be modified. The functions take as input an R data frame with internal SAS names and convert the internal SAS names into the names (ICPSR names) used on the public use files.
6. Using the inputs from steps 3 and 4 and the functions from step 5, a program summarizes to the individual person level the weighted characteristics to be used in the estimation, such as each individual's contribution to the weighted estimate of violent crime or of the prevalence of violent crime for the year. The result is a data frame, together with state

code and the information to be used for variance estimation. Crime variables in the data frame may be summed to check against published values. Similarly, a household-level data frame is created for property crimes.

7. UCR data are extracted and formatted into separate .csv files for each type of UCR crime used as auxiliary predictors in the modeling.
8. R programs read the UCR data from step 7 and data files with information about the NCVS sample design. The scripts then call a function state_model() for each model of interest. For a specified model, state_model() extracts the needed inputs from step 6, fits variance models, and then fits the SAE time-series model itself. The output includes the predicted 3-year averages and their mean square errors. This part of the system requires the most processing time. Researchers may evaluate the fits of alternative models.
9. A simple R script extracts results from the models fitted in step 8 for benchmarking. If alternative models are tried and saved in step 8, this step decides which will be used in the final estimation.
10. Using primarily estimates from the American Community Survey, but also some 2010 census results, a program produces state-level population and housing unit estimates for use in benchmarking.
11. Benchmarking programs read the results from steps 9 and 10. They obtain the national totals for different types of crime by summing the files from step 6 and displaying the totals for checking against the published values. They then use a top-down logic to proportionally adjust the estimated small area rates to be logically consistent. For example, the three categories of violent crime by relationship to the perpetrator are adjusted to sum to the estimated total violent crime rate.

    For three-year periods starting with 2005, the programs form estimated totals of crime by state on the basis of the modeled state rates constrained to match the NCVS population and crime totals. Rounding is incorporated to meet requirements of the Disclosure Review Board, both for the estimated rates and estimated totals. For each characteristic that is benchmarked, separate .csv files are output for the rate, the root mean square error of the rate, and the estimated total for years beginning with 2005-2007.

In general, the above steps are arranged sequentially, but steps 7 and 10 use only publicly available data and can be run when the necessary inputs become available. Step 7 should be revisited for any revision or addition to the UCR state-level data. Step 10 can be performed when the one-year tabulations are released from the American Community Survey, around September in the year following data collection.

To add a new year of data to the system, such as 2019, without adding new variables, steps 1, 3, and 6 need to be completed for the new year. Step 8 can be used to include the new year for production models considered previously and for any new models of interest. Small modifications may be required to the programs at steps 9, 10, and 11.

If new characteristics are of interest, more steps must be rerun. The existing programs for step 1 already extract many variables from the SAS data sets in case they might be of future interest. To analyze one or more of these extracted variables, the functions at step 5 can be checked to ensure that they translate these variables into the ICPSR equivalents, although this is likely to be the case. If not, the functions can be readily modified to add new variables. The program at step 6 needs to be modified to include the new variable or variables. Steps 8, 9, and 11 will then need to be modified accordingly.

To add a new variable or variables that had not been included in the programs for step 1, all of the step 1 programs will need to be modified to add the variable. Steps 2 will need to be repeated to include the new variable, and step 3 will need to be repeated for all past years. Step 4 needs to be rerun, and the functions in step 5 will need to be modified to rename the variable to the ICPSR version. From this point on, the task is essentially the same as in the previous paragraph.

Details of the steps

1. A set of 14 closely related SAS programs extracts data from the Census Bureau's internal SAS data sets from 1998 to 2018. Each one of these programs will extract variables from the SAS data sets for a specific set of one or more years; minor variations in the programs will accommodate changes in the available variables, mostly as possibly useful variables are been included when they appear for the first time. It is possible that the 2017 program will also apply to the 2019 SAS data set, but history suggests that small changes could be needed.

   The large number of variants created to read the SAS files often reflect changes in the names of just a few of the many variables being extracted. A possible way to simplify the maintenance of these programs would be to eliminate some or all of the variables that appear for only a few years if there is little chance they would be modeled. Some variants of the SAS programs would remain necessary, however, to handle changes in spelling of key variables over time.

   The variables used in the current state and substate modeling thus far will be available for all years from 1999 on, but other variables, for example, those used in measuring hate crimes or the edited values for race and Hispanic origin, were added more recently. In general, the variable names on the outputs are taken directly from the SAS data sets, but for 2004 and earlier years spelling differences (e.g., whether the variable name includes "_" or not) are corrected in order to increase the coherence between the .csv data sets for different years.

Separate SAS programs are available:

- extract1998.sas
- extract1999.sas
- extract2000.sas
- extract2002.sas for 2001-2002
- extract2004.sas for 2003-2004
- extract2006.sas for 2005-2006
- extract2009.sas for 2007-2009
- extract2010.sas
- extract2013.sas for 2011-2013
- extract2014.sas
- extract2015.sas
- extract2016.sas
- extract2017.sas for 2017-2018

2. A SAS program extracts information from three internal SAS data sets (one each at the household, person, and incident level) used to create the 2016 bridge data. The files include identifiers and new weights to indicate how to combine the 2015 and original 2016 data sets into the 2016 bridge file. It outputs this information as three .csv files for use in step 3.
3. A short R program, csv_to_Rdata.R, creates R data frames from the .csv files from steps 1 and 2. The outputs are three .csv files at the household, person, and incident levels with the same variable names as the outputs of step 1. This program can be readily modified as needed.
4. A short R program combines the 2015 R data frames, the original 2016 R data frames, and the files from steps 2 and 3 into an R data frame for the 2016 bridge file. The 2017 data frames are also revised with design information used to approximate covariance between the 2017 estimates and the three previous years.
5. Three R functions—hRecode() for households, pRecode() for persons, and iRecode() for incidents—take as input an R data frame with internal SAS names and converts the internal SAS names into the names (ICPSR names) used on the public use files. The functions will accommodate some of the changes in spelling that have occurred in the SAS variable names over time.

   The functions are easily modified to recode additional variables. Each function has two lists; one with the names of extracted variables from the SAS internal files and the other the ICPSR equivalents. Names of new SAS internal variables can be added to the first list and the recoded ICPSR names to the second, carefully preserving the order. The function only recodes variables that it finds present, without reporting an error if it does not find one or more names in the SAS list. Thus, updates to the function do not disturb its ability to handle legacy files.

Generally, all SAS variable names corresponding directly to variables appearing in the ICPSR files will be recoded in this step. Some of the variables for geographic or design information may be recoded but others left with their internal names.

6. A program or sets of programs reads the R data files from steps 3 and 4, calls the functions from step 5, and then aggregates at the individual and household level the contribution to the weighted sum in the numerators and denominators of the incident and prevalence rates for different types of crime. The current version handles both violent and property crime. Each year is processed separately. The outputs are data frames for each year for violent and property crimes with design information and geographic variables. The inclusion of prevalence has resulted in a relatively complex program.

7. The independent variables used in the regression model incorporated in the small area estimation originate from the UCR of the FBI. The method of assembling the UCR data has been updated and greatly simplified to accept results from the new UCR Crime Data Explorer. An R script reformats the reports into .csv files for each time of crime separately. (Temporary corrections were implemented to produce the 2007-2018 estimates, but these will probably not be necessary in the future, as discussed in the next section.) Note that this step and its documentation are not under the constraints of Title 13.

   In the new design, each predictor variable is given in the form of a .csv file with years and columns and geography as rows. The values are in the rates of incidents reported to the police per 100,000, for example, aggrevated.assault.est.csv. In the future, additional predictor variables from other sources could be added by creating files in a similar format and reading them at the beginning of step 8.

8. A program reads the files from step 6 and the reformatted predictors from step 7 and calls a series of functions organized and documented as if they were an R package named NCVSsae2. A primary function, state_model(), manages creation of the rates and covariance matrices for the time-series modeling of state-level estimates, calling other functions in NCVSsae2, functions from the sae2 package and from the survey package. The results of modeling can be saved in .Rdata format for subsequent analysis.

   In the Census Bureau implementation, statements

   ```
   source(“loadnewNCVSsae2package.R”)
   source(“loadnewsae2package.R”)
   ```

are used to load the NCVSsae2 functions and sae2 package with its recent enhancements. The sae2 package available from cran.r-project.org will be updated to agree with the current version used at the Census Bureau.

Different variants of the basic program can be used. For example, the 2007-2018 state estimates were produced using four different versions to cover the different models that were implemented, as discussed in section 8.

Functions included in the NCVSsae2 set specific to the NCVS application have been expanded to model covariances between years 2014-2017, including the covariance resulting from using 2015 data in the 2016 bridge estimates, as well as producing modeled variances and covariances for 2016 and beyond.

9. A simple script extracts results from the models fitted in step 8 for benchmarking.
10. For use in the state-level benchmarking, estimates of the NCVS target population are derived from published tabulations of the ACS and 2010 census. This step is primarily to allow for estimation of state-level totals for victimization as well as rates, as originally requested by BJS for the 1999-2013 state estimates. The estimated state-level totals are available for 2005-2007 and subsequent 3-year periods. Because no title 13 data are involved, the program and its outputs can be shared directly.
11. The 1999-2013 estimates included a final step of benchmarking, applied externally after the estimates from the equivalent of step 7 had been cleared for release by the Control Review Board. The benchmarking programs are now run before submission to the Disclosure Review Board.

UCR State-Level Data

Under the previous small area agreement, state-level estimates were produced for the period 1999-2013 using estimates of crime from the Uniform Crime Reports (UCR) as predictors. For property crime, the NCVS results revealed a logical relationship with UCR equivalents: UCR burglary predicted NCVS burglary, UCR larceny predicted NCVS larceny, and UCR motor vehicle theft predicted NCVS motor vehicle theft. Less consistency was seen been the UCR and the NCVS for violent crime: in particular, UCR aggravated assault was not an effective predictor of any of the NCVS results. UCR robbery was associated with NCVS robbery, and UCR rape appeared associated not only with NCVS combined assault but also with NCVS simple assault. (The UCR does not report the state-level incidence of simple assault.) The previous research documented that the UCR murder rate and the percentage of homeowners from the census did not significantly contribute to the models.

As originally designed in the 1930s, the UCR was based on aggregating summary reports of crime from reporting law enforcement jurisdictions. This original design is now referred to as the Summary Reporting System (SRS), and it is still the approach used in many states. In 1991, the

National Incident-Based Reporting System (NIBRS) was designed to collect detailed information for each criminal occurrence reported to or known to law enforcement. NIBRS data can be aggregated to form totals either equivalent or essentially equivalent to what would be reported by the SRS. Adaption of NIBRS has been slow, with sufficient participation to merit a first annual publication in 2011. By January 1, 2021, however, NIBRS will be the standard for reporting by all jurisdictions. (See the main page for the UCR, https://www.fbi.gov/services/cjis/ucr.)

In preparation to extend the state-level small area models, the beginning of the project surfaced issues meriting closer scrutiny of the UCR data:

1. A previous source for UCR data was eliminated by the FBI, and the possible replacement source, the Crime Data Explorer (CDE) merited careful evaluation, particularly in light of its self-evident errors.
2. A revision in the definition of rape, first implemented in 2013, posed some challenges to using this variable in small area models.

Sources for the UCR estimates

Before the appearance of the FBI’s CDE, https://crime-data-explorer.fr.cloud.gov/, the task of assembling the time series of SRS estimates of crime by state evolved over time and often required more than one source. The annual publication, *Crime in the United States* (CIUS), began in 1958 and remains an archival resource. The annual publications typically included updates of the state estimates for the previous year for some statistics; for example, the publication for 2004 included timely estimates for 2004 and revised estimates for 2003. Beginning in 1995, tables from the publication were available for downloading.

Revision of the SRS state estimates can occur through late reporting by some jurisdictions or by correction of incorrectly reported data. Informally, the impact of the revisions on the small area estimates appears generally modest. In part, the impact of revising UCR estimates for the ending year is blunted by formation of three-year estimates. On balance, it appears better for the sake of timeliness to use the UCR estimates published in the fall in a model using the NCVS data that becomes available at approximately the same time; for example, UCR and NCVS estimates for 2018 were published in the fall of 2019 and used in the current small area estimates for 2007-2018. As a matter of principle, however, it appears best to use revised estimates if they are available at the time.

For a period of time, an online UCR Data Tool periodically reflected revisions in the state-level estimates, but the tool typically presented estimates for the current year months after publication of the CIUS report did. During the previous research for the 1999-2003 estimates, a two-source strategy was employed, using CIUS for the last year in the series when it was unavailable from the UCR Data Tool, and the results from the data tool for other years. The FBI ended support for the data tool, but eventually offered the CDE in its place. A detailed comparison between the CIUS, the data tool, and the CDE was reported in an interim report during the project, which showed that, except for a number of errors in presenting the data, the historical figures from the CDE appeared the best available. Thus, the CDE can be used as the sole source of UCR data for the state-level small area modeling going forward.

In the initial postings, the CDE contained some errors. For example, the CDE initially offered a numeric state code taking the values of the FIPS state codes (e.g., 56 for Wyoming), but these were in error in a number of cases (e.g., codes for Alaska and Alabama were reversed, as were several others). Evident errors were communicated to the email contact address and eventually corrected. For example, the CDE stopped offering a numeric code for states but offered instead the standard two-letter state abbreviation (AK for Alaska, AL for Alabama, etc.). The only apparent error remaining in the CDE estimates used in the current 2007-2018 state estimates was in the 2016 values for rape following the legacy definition. Instead, the values of 2016 rape by the new definition, which were plausible, were proportionately adjusted by 0.723 to approximate the legacy equivalent and used in the modeling for 2007-2018. The error was also communicated to the FBI.

Subsequently, the FBI corrected the 2016 values for rape under the legacy definition. The revised data have a new file name: estimated_crimes_1979_2018_rev1.csv. The estimate of rape under the new definition for 2016 was revised in Hawaii; otherwise, the estimates for all other UCR estimates agreed with the values from the previous file used in modeling the 2007-2018 estimates.

The CDE now appears to be the best source moving forward for UCR/SRS statistics. A short R program, which includes a few built-in checks, has been separately sent to BJS for future use in reformatting the CDE output for use in future modeling.

Revision of the definition of rape in the UCR

The UCR began to implement a revised and more inclusive definition of rape in 2013. That year, the FBI published an addendum to the CIUS accounting for the revision. The legacy definition had only included the rape of females meeting a restricted range of acts, to which the revision added the rape of males, sodomy, and sexual assault with an object. ***Table 1*** of the addendum shows the increase in reported rape to be 41.1%, based on data from NIBRS. The addendum also provided a table-by-table account of which sources were shown in the tables and included in totals for violent crime. The addendum was revised for 2014/2015, and again in 2016. For 2013 through 2016, CIUS published comparisons of the legacy and revised definitions. In jurisdictions covered by NIBRS, both definitions could be implemented on the detailed NIBRS data, but the SRS system collected incidents under only one of the two definitions depending on the jurisdiction, with some jurisdictions switching sooner than others.

Figure 5 shows the relationship between the state estimates under the legacy and revised definitions of rape. After 2013, the relationship is virtually linear, with almost no departures from a straight line. In fact, the state-level estimates may include a mixture of actual comparisons based on the application of the two separate definitions and a linear interpolation of one estimate from the other. The figures for 2013 arguably offer the best comparison of the impact of the definitional change, although fortunately the departures from linearity are not too extreme even in that case. For purposes of future modeling, the program illustrates imputing values for legacy rape after 2016 by multiplying the estimate for rape under the new definition by 0.723.

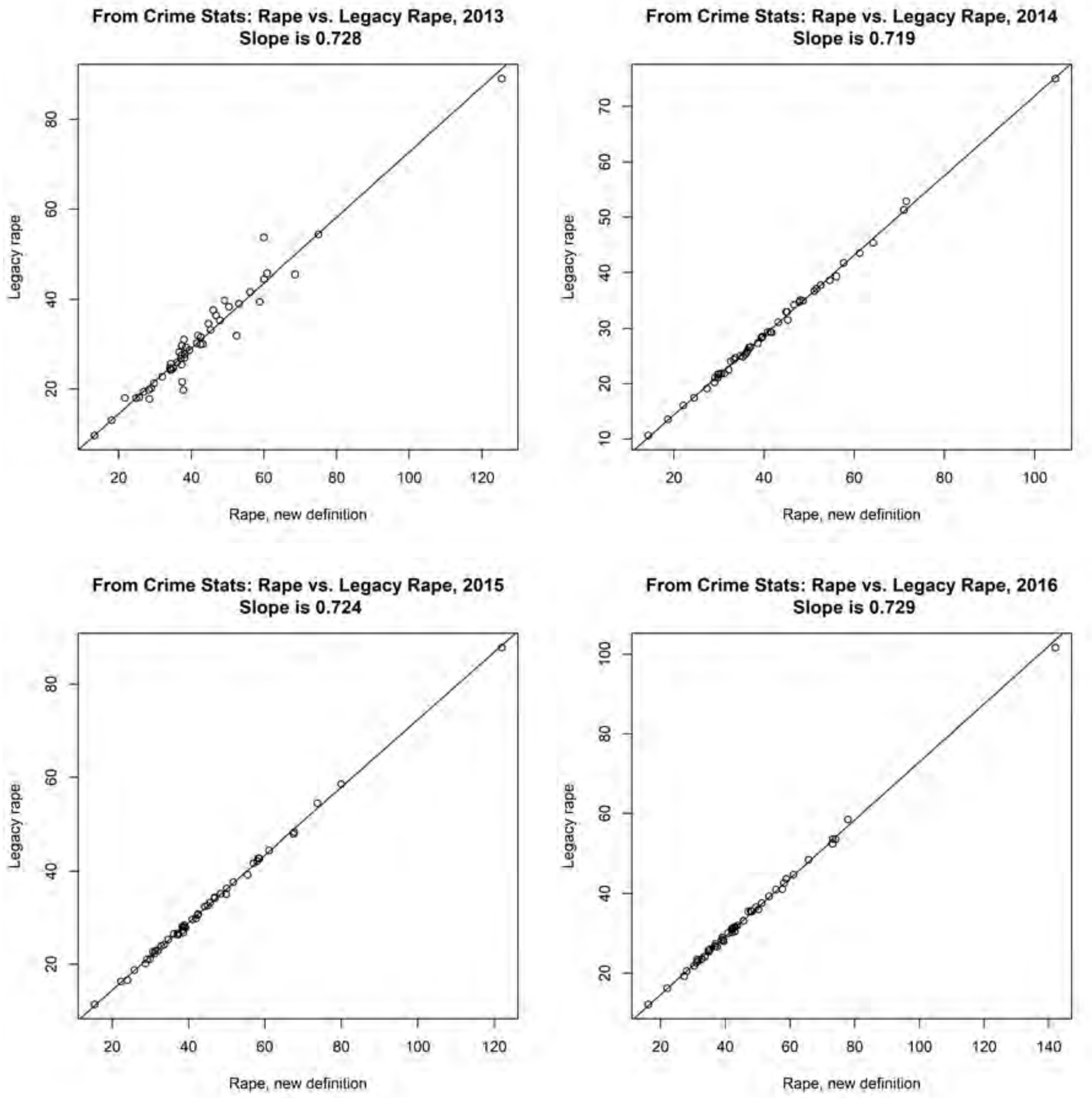

Figure 5. Relationship between the state estimates under the legacy and revised definitions of rape, 2013–2016.

Because the revised definition of rape was first implemented in 2013, a likely path forward would be to apply adjustments to future estimates based on the revised definition to mimic the level of the legacy definition. The estimates in SRS areas are now receiving adjustments of this form, either to make legacy estimates mimic the revised definition or revised estimates mimic the legacy definition. The value of 41.1% reported in the original supplement might be revisited, because the estimates shown in ***Table 1*** of the 2017 CIUS suggest a modest decline, down to 36.0% in 2017. If most of the country is converted to NIBRS in the next few years, there will be a more complete basis to determine what adjustments are necessary, or perhaps even to use the NIBRS data to implement the legacy definition for purposes of the modeling.

Covariance Modeling

The sampling variances of the state direct estimates are a component of the small area model, as are the covariances between estimates for a state of the same characteristic in different years or between different characteristics. (Because a variance is a special case of a covariance, this section will often refer to *covariances* to include *variances*, unless a distinction is explicitly made.) The modeling proceeds by first obtaining direct estimates of covariances at a higher level, such as for self-represent (SR) areas and non-self-representing (NSR) areas, averaging the results, and then distributing the results down to the state level. Section 7.1 details the direct variance calculations, section 7.2 describes the covariance modeling up through 2015, and section 7.3 completes the picture by covering 2016 and after by accounting for the implications of the 2010 sample redesign implemented in 2016.

Direct variance estimates

There are two general approaches to variance estimation for complex sample designs, such as the design of the NCVS: Taylor-series linearization and replication. Both have been applied to NCVS, although linearization has the longer NCVS history. The Census Bureau used linearization to estimate NCVS variances initially, and codes required for the calculations are available on the internal NCVS files over the entire period, beginning in 1997, studied by the small area estimation project. An equivalent pseudo-stratum (V2117) and a secucode(cluster) (V2118) are also available on the NCVS public use files. The small area estimates produced under the first agreement used linearization to produce direct estimates of variance.

More recently, the Census Bureau also supported replication-based variance estimation by producing replicate weights for the survey, and the replicate weights have also been released through ICPSR in recent years. In general, replication provides some advantages over linearization. Replication can more effectively reflect the variance impact of NCVS estimation (weighting) steps on national estimates. But the relative advantage of replication over linearization is far less for direct state and substate estimates. Integration of replicate weighting into NCVS small area estimation system during the second agreement would have posed a number of challenges:

1. The replicate weights are not available for the full span of years under study, necessitating a hybrid approach continuing to rely partially on linearization.
2. The large size of files with the replicate weights was a limiting factor.
3. There was no coordination between the replicate weights for 2015 with the replicate weights designed for 2016 and after, which would require a specialized solution to derive covariances between the two time periods.

Consequently, the research under the second agreement continued by extending the linearization approach used during the first agreement.

To capture much of the covariance between the estimates up to 2015 and those for 2016 and after, the pseudo-stratum code was modified for the small area variance calculations beginning in 2016. All 2015 data used in the 2016 estimates retained their 2015 pseudo-stratum assignments. In 2016 and 2017, any sample in self-representing areas that was also in sample in 2015 was reassigned the same pseudo-stratum and secucodes as assigned in 2015. Otherwise, the variance codes were modified to ensure that they would not overlap with those in 2015 or earlier years. For sample in the 22 supplemented states, a value equal to 1000 times the FIPS state code was added to the pseudo-stratum code, thus strictly reflecting the state stratification in the variance calculation. For the 2010 based sample in the remaining states, 500 was added to the pseudo-stratum code, which was sufficient to distinguish these cases from any in 2015 or earlier years. The non-supplemented states were also thus clearly separated from the supplemented states.

Direct variance estimation was based on the observation that every dependent variable included in an NCVS small area model is a ratio of two estimated totals. For example, the violent crime rate is the ratio of the estimated total of violent crime incidents to the estimated total number of persons. For some of the ratios studied in the project, the numerator and denominator use different weights, but they share the same sample design. A standard approach to linearization for ratios, deriving linear substitutes, was used to produce direct variance and covariance estimates for ratios.

Modeling covariances over 1997-2015

Although supported by standard theory, the direct variance estimates are unusable at the state level for the small area modeling. For example, there could easily be no reports of a specific type of crime in a given state and year, but the resulting direct variance estimate of zero is inadequate. Consequently, modeling the variances and covariances was required for the NCVS applications.

As noted in Section 2, the state estimates for 1999-2013 produced during the first agreement depended on the sampling covariances (including variances) of the direct state-level estimates. For modeling a single characteristic, such as total violent crime, the required covariances were across years within each state, with covariances between states treated as zero. For multivariate models of two or three characteristics, the required covariances included those for each characteristic separately and also those between pairs of characteristics within the state over time, but with covariances between states again treated as zero. In other words, when the estimates are ordered grouped by state, the corresponding covariance matrix was assumed to be block diagonal.

During the first agreement, research on the existing NCVS design identified between-PSU variance in NSR areas as a major contributor to the total sampling variance. The research also found that correlations over time were quite different in NSR areas compared to SR areas: NSR areas exhibited both much larger design effects and correlations that more slowly dampened out over time. Between sample redesigns, the NSR design remained in the same sampled PSUs, contributing to the larger correlation over time.

The 2015 final report (Fay and Diallo, 2015c) summarized the modeling of the covariances, which began with a separate direct estimation of covariances for years falling in the intervals 1997-2004 and 2005-2015 for SR and NSR areas separately. For example, the years were 1999-2004 and 2005-2013 for the 1999-2013 estimates. Three modeling steps followed:

1. The directly estimated SR and NSR covariance matrices were converted to correlation matrices, which were then separately modeled and smoothed versions produced.
2. The smoothed correlation matrices were converted into smoothed covariance matrices based on the directly estimated SR and NSR variances for each year.
3. For any given state, its modeled estimate of the covariance combined the SR and NSR smoothed covariances in relation to the state's expected sample sizes in SR and NSR areas.

Under the first agreement, the approach adopted to model the SR and NSR covariance matrices for the period 1997-2015 continued to be used for this span of years during the current agreement. Directly estimated covariances for the SR and NSR portions of the sample are first estimated for the appropriate years (e.g., 1999-2004 and 2005-2013). Because of the substantial complexity of the 2000 census redesign in 2005, covariances between 1999-2004 and 2005-2013 are set to 0. And because the NCVS sample size varies from one year to the next, the sampling variances (that is, the diagonal elements of the full covariance matrix) are not modeled. But, given the general stability of the design during this period, the correlations between years are averaged depending on the difference in years for the pair of estimates.

For example, an average correlation is computed for estimates of the same characteristic one year apart, excluding the 2004/2005 boundary. Year 2005 is also excluded from the averaging because its design information on the internal file is not consistent with 2006 and subsequent years. For 1999-2013, the following 11 pairs are included in the average: 1999/2000, 2000/2001, 2001/2002, 2002/2003, 2003/2004, 2006/2007, 2007/2008, 2009/2010, 2010/2011, 2011/2012, and 2012/2013. For a 2-year difference, 9 pairs are averaged: 1999/2001, 2000/2002, 2001/2003, 2002/2004, 2006/2008, 2007/2009, 2009/2011, 2010/2012, and 2011/2013. The pair 2006/2013 provides the only instance of a 7-year difference in this case.

After average correlations are obtained for each difference in years, they are further smoothed by fitting a linear regression to them, weighted by the number of paired years used in the calculation of each difference. In other words, differences of one year are given the most weight because they are supported by the most data, and the largest differences with only a few or one observation 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. Correlations between years 1997-2004 and 2005-2015 remained zero.

When two or three characteristics are being modeled, the covariance between them in the same year is kept fixed while 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 one 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.

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.

After modeled covariances C_{SR} and C_{NSR} for SR and NSR areas have been obtained, they are combined to form an estimate of the state-level covariance. 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} the total SR and NSR sample sizes. Letting

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

the modeled state-level covariance is

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

For the period 1997-2015, the actual SR and NSR sample size in each state is not used because the national NCVS sample design was not stratified by state, and the sample size in the NSR areas of a state was random and typically highly variable. A small state with much of its population in NSR areas might not include any of its NSR PSUs in the sample; therefore, simply using the observed NSR sample size as zero could substantially understate the uncertainty in the direct state estimate in this situation. Instead, the approach taken is to attempt to approximate the unconditional variance based on the expected sample size in SR and NSR areas under the sample design.

Previous work by Fay and Li (2012) provided estimates of the expected sample sizes in SR and NSR areas. The original purpose was to estimate the extent to which the NCVS might support direct state-level estimates if the sample size were supplemented. Rather than use confidential information about the design, they constructed a "public model" of the sample design that could be constructed given publicly available information. It was clear from NCVS documentation that the survey included the largest metropolitan areas as SR. Although the exact SR/NSR boundary was not published, the model was based on a reasoned guess for the size of this boundary based on interviewer workloads. The public model was intended to predict the approximate location of the SR sample but not which NSR PSUs had been selected; thus it did not pose a threat to the confidentiality of the NCVS sample. (As a side note, by building a public model of how the sample design would evolve when a state was supplemented and more NSR areas became SR, the authors offered estimates of how much supplementation would be required to produce a c.v. of 10% on a 2% violent crime rate for different numbers of supplemented states. The model and its further elaboration helped to support the subsequent decision to supplement a total of 22 states, beginning in 2016.)

Modeling covariances over 2014-2018 and after

The state supplementation introduced in 2016 posed both an opportunity to use the increased sample size to advantage and a challenge to properly model covariances for a design that was no longer approximately self-weighting.

To mitigate the variation in weights starting in 2016, states were divided into two groups:

1. Group 1: 18 of the 22 supplemented states, excluding California, Florida, New York, and Texas.
2. Group 2: California, Florida, New York, Texas, the non-supplemented states, and the District of Columbia.

Average weights in Group 1 were substantially smaller than in Group 2. Weight variation within each group was relatively much less.

Depending on the range of years specified for the small area model, the current system computes four sets of direct covariance matrices:

1. SR and NSR covariances for years in 1997-2004,
2. SR and NSR covariances for years in 2005-2015,
3. SR and NSR covariances for years in 2014-2018,
4. SR and NSR covariances separately for Group 1 and Group 2 for years 2016 and beyond.

The first two sets of covariance matrices are modeled just as under the first agreement, up to year 2015, as described in the previous section.

The treatment of set 4 is more straightforward than the third, so it will be described first. In short, the model for the fourth set is similar to sets 1 and 2, but it is implemented for Group 1 and Group 2 states separately. The averaging of correlations will not begin until 2016-2019; until then the directly estimated SR and NSR covariances for 2016-2018 will be 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.

The method of combining the SR and NSR covariances is modified for set 4. Because stratification is within state, a conditional analysis based on the realized sample size does not present the same issues associated with the national design in 1997-2015. For set 4, the SR proportion of the state population is estimated on a weighted basis,

$$f_i = \frac{\hat{N}_{SR,i}}{\hat{N}_{SR,i}+\hat{N}_{NSR,i}},$$

where $\hat{N}_{SR,i}$ and $\hat{N}_{NSR,i}$ 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

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

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

The set 3 covariances calculated over 2014-2018 are then used to estimate correlations of estimates in 2014 or 2015 with estimates in 2016, 2017, or 2018. Direct SR and NSR covariances are estimated using the design information without dividing into state groups as for set 4. The resulting covariance matrices are combined with the matrices from sets 1 and 2 to produce a matrix up to year 2018 based on the 2000 design, but continuing to use the smoothed covariances based on set 2 up to 2015.

The SR and NSR covariances from sets 1, 2, and 3 are combined according to the formulas originally used for sets 1 and 2 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 by:

1. Any variance or covariance between estimates for 2016 and beyond uses the results from set 4.
2. A covariance between an estimate in the range 2016-2018 and one in 2014-2015 is adjusted to reflect the difference between the variance estimated for the 2016-2018 estimate 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 set 1-3 variance.

This approach produces a covariance estimate based on the correlation from set 3 and the variance from set 4.

Software implementation

Section 5 introduced the steps of the small area system, and communication of information about the design is included in several of them:

- In step 1, the extract programs capture UCF_PSEUDOSTR and UCF_HALFSAMPCD, as well as geographic information
- In step 3, the program csv_to_Rdata_SAS.R recodes UCF_PSEUDOSTR as described in section 7.1 for both supplemented and unsupplemented states, beginning in 2018.
- The program create2016bridge.R combines the original 2016 data and 2015 data along with the prescribed weights for the 2016 bridge file, but does not alter UCF_PSEUDOSTR or UCF_HALFSAMPCD.
- Another program, reviseVar.R, reads in the 2015 and preliminary 2016 and 2017 data sets for households and alters the 2016 and 2017 variance codes as described in section 7.1 after determining which households were in the 2015 sample. This program is an intermediate step between steps 4 and 5.
- In step 8, the function state_model() in the NCVSsae2 package orchestrates the covariance estimation. It calls geo_ratios() in the sae2 package to create the linear substitutes for each year corresponding to the ratios being estimated, separately for SR and NSR areas. If set 4 is to be computed, state_model() again calls geo_ratios() to create 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, 2, and 3 and to vcov_state_sup() for set 4. In turn, vcov_state() and vcov_state_sup() call vcovgen() in the sae2 package, which relies on svytotal() and vcov() in the survey package for the actual calculations. Both vcov_state() and vcov_state_sup() (when 2019 is included) call smooth_cov(), which in turn calls corr_average() and expand_corr(). The function corr_average() calls rbar_calc(), rbar_cross(), row_smooth(), and r_count().

Modeling the 2007-2018 Estimates

Section 5 summarized the software system for producing the small area estimates as a series of steps. Step 8 integrated inputs from the previous steps for use in producing small area models leading to the final small area estimates. Step 9 followed with the role of selecting specific results to finalize. This section will discuss the R programs that produced the 2007-2018 state estimates, both to document the models that were selected and to illustrate how this step can produce estimates for other periods or other characteristics. The programs have been modified only to remove information about the specific location of the files at the Census Bureau, in order to protect information that the Census Bureau deems confidential. In this form, the files from steps 8 and 9 have been transmitted to BJS.

The four programs for step 8 are named fits22Oct2019v2018_extract.R, fits25Oct2019v2018_extract.R, fits26Oct2019v2018_extract.R, and fits08Nov2019v2018_extract.R. The names are of course arbitrary, but they reflect the approximate date of the first version, the use of UCR data released with final year 2018, and the fact that details of the file locations have been removed. Each starts with a sequence of statements loading information to be used, followed by a sequence of model statements in arbitrary order specifying models to be fitted, and ending with a statement to save the models. Sometimes, an intermediate save was inserted in order to allow segments of the program to be run on separate occasions. The results would have been equivalent if a single long program was run with all of the models, but using four programs instead facilitated management of the model fitting.

Two statements at the beginning of these programs determine the span of years to be fitted.

```
pred.start <- 2007
pred.end  <- 2018
```

By changing these two statements to indicate a different span of years, the rest of the program would run accordingly, as long as the other statements in the first section identify files include the information for this range of years. A change to

```
pred.start <- 1997
pred.end  <- 2003
```

would fit the models on the original span of years, without requiring any other change. Thus, input files only need to grow by incorporating new data as it becomes available, and the program will select the years needed for the modeling.

The models for the incidence of property crime are simpler than the others:

```
system.time(tot.prop.2018 <- state_model(                                    # (5)
    total.prop ~ ucr.larceny + ucr.auto.theft + ucr.burglary + year + 0,
    denominators="WGTHHCY",

    pred.start=pred.start, pred.end=pred.end,
    UCR=UCR, suff="prop",
    census2010=census2010, modeled.design=modeled.design,
    patch="state_model_patch.R",
    file.list=file.list.prop, detail=TRUE))

system.time(prop2.2018 <- state_model(list(                                   # (6)
    burglary ~ ucr.burglary + year + 0,
    comb.theft ~ ucr.comb.theft + year + 0),

    pred.start=pred.start, pred.end=pred.end,
    denominators="WGTHHCY",
    UCR=UCR, suff="prop",
    census2010=census2010, modeled.design=modeled.design,
    patch="state_model_patch.R",
    file.list=file.list.prop, detail=TRUE))

system.time(prop2b.2018 <- state_model(list(                                        # (7)
    auto.theft ~ ucr.auto.theft + year + 0,
    larceny ~ ucr.larceny + year + 0),
    pred.start=pred.start, pred.end=pred.end,
    denominators="WGTHHCY",
    UCR=UCR, suff="prop",
    census2010=census2010, modeled.design=modeled.design,
    patch="state_model_patch.R",
    file.list=file.list.prop, detail=TRUE))
```

Three models are fitted, which were stored with names tot.prop.2018, prop2.2018, and prop2b.2018. The first model uses UCR larceny, motor vehicle theft, and burglary as fixed predictors in the regression model. Similarly, a bi-variate model predicts burglary on the basis of UCR burglary and combined theft on the basis of the UCR analogue. A second bi-variate model predicts auto theft and larceny from their UCR equivalents. This strategy of modeling separated burglary as a somewhat distinct characteristic. Other approaches, including modeling all three components of property crime simultaneously, were possible. This modeling approach had been used in 1999-2013 to dampen the impact of a possible outlier on the coefficient for motor vehicle theft, and it was continued into the current application.

Benchmarking

The issue of benchmarking arises in many small area applications when models for different levels of aggregation produce estimates that violate, even if usually only slightly, logical relationships that should hold exactly. In the NCVS application, for example, estimates of the incidence of violent crime by type should logically add up to the incidence of total violent crime, but the small area model for crime by type produces estimates differing somewhat from the univariate model for total violent crime. During the first agreement, a bottom-up approach was considered, where the estimates of total violent crime would be the sum of the estimates by type of crime. Not surprisingly, however, the results from this approach disagreed with an alternative bottom-up approach by estimating violent crime by relationship to the perpetrator. There was no convincing basis to choose between these two bottom-up alternatives for violent crime. Furthermore, the univariate model for total violent crime seemed to give more stable results than either bottom-up alternative, presumably because the univariate model was based on more data and estimated fewer parameters.

Near the end of the first agreement, a top-down approach was instead adopted and implemented for the final estimates, both at the state and substate levels. The top-down approach anchored the estimates to univariate models for total violent crime and total property crime, and the approach constrained estimates for their components to agree to the totals through benchmarking. Section 9.1 describes the simple approach taken to benchmark rates or proportions for large counties and CBSAs, and for state estimates before the 2005-2007 period.

During the first agreement, BJS also requested that estimated numbers of victimizations be produced, and by agreement this request was met for states by multiplying the small area estimated rates by estimates of the NCVS-eligible population based on the American Community Survey (ACS) and 2010 Census, beginning with 2005-2007 when estimates from the ACS became available. Section 9.2 describes the estimation of the eligible NCVS population based on published data from these two sources. Section 9.3 then describes benchmarking for incidence and section 9.4, for prevalence, when estimated numbers of victimizations (for incidence) or victims (for prevalence) are also produced. Section 9.5 describes the software implementation of the benchmarking and production of the population estimates.

Benchmarking of rates and proportions

The top-down approach adopted to benchmark incidence rates is relatively straightforward. Suppose a total for a major class of crime (violent crime or property crime), t, is broken down into three components, a, b, and c. If $\hat{p}_{t,i}$ represents the SAE estimate of the incidence rate of t in state i in a given year from a univariate model, and if $\hat{p}_{a,i}$, $\hat{p}_{b,i}$, and $\hat{p}_{c,i}$ are estimates of the components either from a multivariate model or three univariate models, then the components can be each multiplied by the factor $\hat{p}_{t,i}/(\hat{p}_{a,i} + \hat{p}_{b,i} + \hat{p}_{c,i})$ so that the sum of the components will be consistent with the estimated total.

As an alternative to modeling the three components jointly, some models dichotomize the total by combining two of the components, modeling the resulting two components, and then separately modeling the original two components that had been combined at the first step. For example, if b and c are combined into bc and modeled jointly with a, the resulting two modeled components are adjusted by factor $f_{1,i} = \hat{p}_{t,i}/(\hat{p}_{a,i} + \hat{p}_{bc,i})$, giving the adjusted estimates for a. When b and c are modeled separately, their resulting predictions are adjusted by $f_{2,i} = f_{1,i}\hat{p}_{bc,i}/(\hat{p}_{b,i} + \hat{p}_{c,i})$. For the 2007-2018 state estimates, total violent crime is first dichotomized into (1) violent crime except simple assault and (2) simple assault, and these two components are modeled. The next model is of the dichotomy (1) aggravated assault/rape and (2) robbery. Similarly, property crime is dichotomized into (1) burglary and (2) combined theft, and combined theft is dichotomized into (1) motor vehicle theft and (2) larceny.

Benchmarking of prevalence proportions is somewhat more complex because an inequality rather than an equality is involved. The sum of prevalence proportions for components a, b, and c can exceed the prevalence of t, but it cannot be less. After modeling the prevalence of the total, $\hat{p}_{t,i}$, state-level estimates of the sum of the component prevalence proportions, $\hat{p}_{s,i}$ are required. To attempt to impose the logical relationship $\hat{p}_{s,i} \geq \hat{p}_{t,i}$, the ratio $r_i = p_{s,i}/p_{t,i}$ is modeled based on observed ratios, and then the estimate $\hat{p}_{s,i} = \hat{r}_i \hat{p}_{t,i}$ is formed. The adjustment factor for the prevalence rates for the components is then $\hat{p}_{s,i}/(\hat{p}_{a,i} + \hat{p}_{b,i} + \hat{p}_{c,i})$.

A similar logic is applied when a total is dichotomized by combining two of the initial categories. In this case, the sum of the prevalence proportions, $\hat{p}_{s,i}$, combines the prevalence of a with the prevalence of bc, that is, the prevalence proportion of the combined bc category. The estimation may again be approached by modeling the ratio, $\hat{r}_{1,i}$, between the sum of these two prevalence proportions and the total prevalence. The needed sum is then estimated by $\hat{p}_{1s,i} = \hat{r}_{1,i} \hat{p}_{t,i}$ and used to compute $f_{1,i} = \hat{p}_{1s,i}/(\hat{p}_{a,i} + \hat{p}_{bc,i})$. When b and c are then separately modeled, the sum of the prevalence proportions of b and c is compared to the prevalence of the combined bc category. The resulting modeled ratio, $\hat{r}_{2,i}$, is then used in $f_{2,i} = f_{1,i} \hat{r}_{2,i}\, \hat{p}_{bc,i}/(\hat{p}_{b,i} + \hat{p}_{c,i})$ to adjust $\hat{p}_b$ and $\hat{p}_c$.

Population and household controls from the ACS

Although the small area models produced estimates for single years, they are summarized for publication in the form of 3-year averages. To meet the request to produce estimated totals at the state level consistent with the estimated rates, 3-year averages of population were required. Originally, the ACS produced 3-year averages as well as single years and 5-year averages, but when ACS publications were reduced to single years and to 5-year averages in 2014 (that is, 2011-2013 is the last published set of 3-year averages), 3-year averages have been formed from the single-year data.

Essentially, the universe for the NCVS is the non-institutional population age 12+. The ACS provides estimates for the total population and the population in group quarters, but the fraction, r_i, of the group quarters population in institutions in state i must be estimated from published 2010 census results. From the ACS estimate of the total population, three quantities were subtracted:

1. The ACS group quarters population $\times\ r_i$
2. The ACS population age 0-11 in households
3. The ACS population under age 18 in group quarters $\times\ (1 - r_i)$

The third term attempts to approximate the population under age 12 in non-institutional group quarters. The age difference is an obvious imperfection, but further detail was unavailable and the component was quite small.

The ACS similarly provides estimated numbers of households used for benchmarking property crime. Further detail on the computer implementation is included in section 9.5.

Benchmarking of incident rates and estimated numbers

The numbers of victimizations by state were estimated by multiplying the small area estimate of the rate by the estimated population. The method of benchmarking required modification to avoid producing estimated numbers inconsistent with NCVS national totals. The following steps were used to benchmark the 2007-2018 state estimates:

1. Proportionally adjust the ACS-based population estimates by state from section 9.2 to agree with the 3-year averages of the NCVS population.
2. Compute initial estimated numbers of victimizations by violent crime by state, by multiplying the population estimates from step 1 by the small area estimates of violent crime.
3. Proportionally adjust the initial estimates from step 2 to agree with the NCVS 3-year average estimate of total violent crime.
4. The models for violent crime by relationship to the perpetrator are first based on the dichotomy between crime by strangers and all non-strangers. Compute initial estimated numbers of victimizations by state from the small area modeled rates and step 1. Then use 2-dimensional raking (iterative proportional fitting) to adjust the preliminary estimates to agree to the NCVS national estimates of violent crime by strangers and crime by non-strangers (the first margin) and the state-level estimates from step 3 (the second margin).
5. The second model for relationship to the perpetrator is intimate partner violence compared to all other non-strangers. Again, compute initial estimated numbers and then rake these to the national estimates for these two categories and the state-level estimates for all non-stranger violence from step 4.
6. The models for violent crime by type are based on the dichotomy between simple assault and violent crime excluding assault (called "serious violent crime" in publications before 2018). Compute initial estimated numbers of victimizations by state from the small area modeled rates and step 1. Then use 2-dimensional raking to adjust the preliminary estimates to agree to the NCVS national estimates of simple assault and violent crime excluding simple assault.
7. The second model for crime by type is for the dichotomy (1) robbery and (2) aggravated assault/rape. Again, compute initial estimated numbers and then rake these to the national estimates for these two categories and the state-level estimates for violent crime excluding simple assault from step 6.

8. Estimates from the preceding five steps are converted to rates per 1,000 based on the population estimates from step 1 and rounded.
9. Root mean square errors are also converted to rates per 1,000 and rounded, based on the results from the model used in estimating each component. The RMSE estimates are not adjusted for the effect of the raking, which would require additional methodological development.

Benchmarking the incidence of property crime for 2007-2018 is similar:

1. Proportionally adjust the ACS-based household estimates by state from section 9.2 to agree with the 3-year averages of the NCVS estimated households.
2. Compute initial estimated numbers of property crime victimizations by state by multiplying the household estimates from step 1 by the small area estimates of total property crime.
3. Proportionally adjust the initial estimates from step 2 to agree with the NCVS 3-year average estimate of total property crime.
4. The models for property crime by type are first based on the dichotomy between burglary and total theft including motor vehicle theft. Compute initial estimated numbers of property crime victimizations by state from the small area modeled rates and step 1. Then use 2-dimensional raking to adjust the preliminary estimates to agree to the NCVS national estimates of burglary and total theft and the state-level estimates from step 3.
5. The second dichotomy model for property crime is motor vehicle theft compared to other theft. Again, compute initial estimated numbers and then rake these to the national estimates for these two categories and the state-level estimates for total theft from step 4.
6. Estimates from the preceding four steps are converted to rates per 1,000 based on the population estimates from step 1 and rounded.
7. Root mean square errors are also converted to rates per 1,000 and rounded, based on the results from the model used in estimating each component.

Benchmarking of prevalence proportions and estimated numbers

Prevalence numbers are similarly estimated by multiplying the small area estimate of the proportion by the estimated population. The principles for benchmarking prevalence proportions are generally similar to those for incidence estimates.

1. As in 9.3, proportionally adjust the ACS-based population estimates by state from section 9.2 to agree with the 3-year averages of the NCVS population.
2. Compute initial estimated numbers of victims of violent crime by state, by multiplying the population estimates from step 1 by the small area estimates of the prevalence of violent crime.

3. Proportionally adjust the initial estimates from step 2 to agree with the NCVS 3-year average estimate of the number of victims of violent crime.
4. Form an initial estimate of the sum of the prevalence estimates by state for the three categories of relationship to the perpetrator as the product of the small area estimates of the prevalence of violent crime times the estimate from a univariate model of the ratio of the sum of the three prevalences to the prevalence of violent crime.
5. Proportionally adjust the preliminary state-level estimates from step 4 to agree with the national sum of the three prevalences by relationship.
6. Rake the state-level prevalences for stranger, intimate partner, and other non-stranger to agree with the national totals for each of the three categories and the state-level results from step 5.
7. Form an initial estimate of the sum of the prevalences of simple assault and of violent crime excluding simple assault as the product of the small area estimates of the prevalence of violent crime time times the estimate from a multivariate model of the ratio of the sum of the two prevalences to the prevalence of violent crime. (The prevalence of violent crime was modeled simultaneously with the ratio.)
8. Proportionally adjust the preliminary state-level estimates from step 7 to agree with the national sum of the two prevalences by type of crime.
9. Rake the state-level prevalences for simple assault and for violent crime excluding simple assault to agree with the national totals for the two categories and the state-level results from step 8.
10. Form an initial estimate of the sum of the prevalences of aggravated assault/rape and of robbery as the product of the small area estimates of the prevalence of violent crime excluding assault times the estimate from a multivariate model of the ratio of the sum of the two prevalences to the prevalence of violent crime excluding assault. (The prevalence of violent crime excluding assault was modeled simultaneously with the ratio.)
11. Proportionally adjust the preliminary state-level estimates from step 10 to agree with the national sum of the two prevalences by type of crime.
12. Rake the state-level prevalences for aggravated assault/rape and for robbery to agree with the national totals for the two categories and the state-level results from step 11.
13. Convert estimates from steps 3, 6, 9, and 12 to rates per 1,000 based on the population estimates from step 1 and round.
14. Convert root mean square errors to rates per 1,000 and round, based on the results from the model used in estimating each component.

Note that the decision to use a univariate model for the ratios in step 4 was based on difficulty modeling a multivariate model with total violent crime as the other component. In steps 7 and 10 ratios were modeled jointly with their denominators.

Benchmarking the prevalence of property crime involves similar principles.

1. As in 9.3, proportionally adjust the ACS-based household estimates by state from section 9.2 to agree with the 3-year averages of the NCVS estimated households.
2. Compute initial estimated numbers of victimized households of property crime by state, by multiplying the household estimates from step 1 by the small area estimates of the prevalence of property crime.
3. Proportionally adjust the initial estimates from step 2 to agree with the NCVS 3-year average estimate of the number of households victimized by property crime.
4. Form an initial estimate of the sum of the prevalence estimates by state for the three categories of property crime—burglary, motor-vehicle theft, and other theft—as the product of the small area estimates of the prevalence of property crime times the estimate from a multivariate model of the ratio of the sum of the three prevalences to the prevalence of property crime. (The prevalence of property crime was modeled simultaneously with the ratio.)
5. Proportionally adjust the preliminary state-level estimates from step 4 to agree with the national sum of the three prevalences by type of property crime.
6. Rake the state-level prevalences for burglary, intimate partner, and other non-stranger to agree with the national totals for each of the three categories and the state-level results from step 5.
7. Convert estimates from steps 3 and 6 to rates per 1,000 based on the population estimates from step 1 and round.
8. Convert root mean square errors to rates per 1,000 and round, based on the results from the model used in estimating each component.

Software implementation

The R scripts state_pop_controls_2007_2018.R and state_hh_controls_2007_2018.R read downloaded tabulations from the 2010 Census and ACS and compute the 3-year averages as described in section 9.2. Specifically, the population version reads tables PCO1 and PCO2 from the 2010 Census and tables B01003, B09001, and B26001 from the ACS, and the household version reads table B25002. Much of the code is designed to accommodate both the changing formats used by the ACS and the switch from using published 3-year averages to combining 1-year estimates into 3-year averages. The two scripts each output a single file for use the benchmarking either violent or property crime estimates.

The code accommodates the new format for the 2018 tables introduced for the move from American Fact Finder to data.census.gov. The 2019 ACS single-year data are expected in the fall of 2020. If they are published in the same format as 2018, then a code change in the scripts of “pred.end <- 2018” to “pred.end <- 2019” and “if (iyear == 2018)“ to “if (iyear >= 2018)” will update the script to include 2019. In general, extending the scripts for future years will require reviewing new data against old to determine if any new changes in format have occurred.

Before benchmarking, previous scripts may have called state_model() with competing models and the results saved for consideration, which was the case in fitting the state models for 2007-2018. Before benchmarking itself, two scripts for the 2007-2018 state estimates, extract3year2018ser_extract.R (for incidence) and extract3year2018prevser_extract.R (for prevalence) select and extract the specific results to be included in the final estimates. The two scripts have a simple structure of loading the files with the results from state_model(), creating variables to be recognized by the benchmarking program, extracting the specific contents for the small area estimates of the 3-year averages and their root mean square errors, and outputting the results in a convenient form for use by the benchmarking programs.

The benchmarking R scripts are more complex:

- benchmark_viol_ser_2018_extract.R, for incidence of violent crime
- benchmark_viol_prev_ser_2018_extract.R, for prevalence of violent crime
- benchmark_prop_cen_2018_extract.R, for incidence of property crime
- benchmark_prop_prev_cen_2018_extract.R, for prevalence of property crime

Each 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, and the scripts are written generally to allow changing these values within the same range as state_model(), that is pred_start is 1997 or larger, but not equal to 2006, and pred_end is 2007 or greater. NCVS national totals are aggregated from the same files used to fit the models and 3-year estimates. The scripts include the special formations of the 2004-2006, 2005-2007, and 2006-2008 averages by excluding 2006 results and averaging the other two years.

The benchmarking scripts each include identical functions ncvs_adjust_1(), rake.2(), ncvs_adjust_2(), ncvs_adjust_3() that generalize the adjustments described in sections 9.1, 9.3, and 9.4. Unlike the set of more formal functions in NCVSsae2 or the sae2 package, these functions reference some values from the global environment rather than requiring all needed values to be in their argument list. (From the perspective of documentation, it may be helpful to keep these functions imbedded in the context of their use within the benchmarking scripts.)

The scripts otherwise follow the order of operations outlined in sections 9.1, 9.3, and 9.4. The outputs are three .csv files for each publication variable, with rates or proportions stated per 1,000 and rounded to the same degree as the publication of the 1997-2013 state estimates.

References

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

Cantor, D., Krentke, T., Stukel, D., and Rizzo, L. (2010). NCVS Task 4 Report: Summary of Options Relating to Local Area Estimation. Submitted by Westat to the Bureau of Justice Statistics, May 19, 2010, http://www.bjs.gov/content/pub/pdf/westat_lae_5-19-10.pdf.

Catalano, S. (2012). Intimate Partner Violence, 1993-2010. Bureau of Justice Statistics. NCJ 239203, November 2012.

Cork, D.L and Groves, R.M (eds.) (2008). Surveying Victims: Options for Conducting the National Crime Victimization Survey. National Academies Press, Washington, DC.

Fay, R.E. and Diallo, M. (2015a). Developmental Estimates of Subnational Crime Rates Based on the National Crime Victimization Survey. Bureau of Justice Statistics. NCJ 249238, December 2015. BJS web, https://www.bjs.gov/index.cfm?ty=pbdetail&iid=5499.

Fay, R.E. and Diallo, M. (2015b). Developmental Estimates of Subnational Crime Rates Based on the National Crime Victimization Survey, 1999-2013: Summary Report. Westat, Inc. Unpublished report dated March 24, 2015.

Fay, R.E. and Diallo, M. (2015c). Constructing and Disseminating Small Area Estimates for the National Crime Victimization Survey (NCVS): Continuation of Project 2008-BJ-CX-K067. Westat, Inc. Unpublished report dated March 30, 2015.

Fay, R.E. and Li, J. (2012). Rethinking the NCVS: Subnational Goals through Direct Estimation. FCSM Research Conference. https://nces.ed.gov/FCSM/2012research.asp#TuesdayAM.

Lauritsen, J.L., Owens, J.G, Planty, M., Rand, M.R., and Truman, J.L. (2012). Methods for Counting High-Frequency Repeat Victimizations in the National Crime Victimization Survey. Bureau of Justice Statistics. NCJ 237308, April 2012.

Lynch, J.P. and Addington, L.A. (eds.) (2007). Understanding Crime Statistics: Revisiting the Divergence of the NCVS. Cambridge University Press.

Morgan, R.E. and Kena, G. (2017). Criminal Victimization, 2016. Bureau of Justice Statistics. NCJ 251150, December, 2017. BJS web, http://www.bjs.gov/content/pub/pdf/cv16_old.pdf.

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

Morgan, R.E. and Oudekerk, B.A. (2019). Criminal Victimization, 2018. Bureau of Justice Statistics. NCJ 253043, September, 2019. BJS web, http://www.bjs.gov/content/pub/pdf/cv18.pdf.

Morgan, R.E. and Truman, J.L. (2018). Criminal Victimization, 2017. Bureau of Justice Statistics. NCJ 252472, December, 2018. BJS web, http://www.bjs.gov/content/pub/pdf/cv17.pdf.

Rand, M. and Cantalano, S. (2007). Criminal Victimization, 2006. Bureau of Justice Statistics. NCJ 219413, December, 2007. BJS web, http://www.bjs.gov/content/pub/pdf/cv17.pdf.

Truman, J.L. and Langton, L. (2014). Criminal Victimization, 2013. Bureau of Justice Statistics. NCJ 247648, Sept. 18, 2014. BJS web, http://www.bjs.gov/content/pub/pdf/cv13.pdf.