Estimating County-Level Aggravated Assault Rates by Combining Data from the National Crime Victimization Survey (NCVS) and the National Incident-Based Reporting System (NIBRS)

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Elizabeth Ellen Petraglia, M.S.
Graduate Program in Statistics

The Ohio State University

2015

Dissertation Committee:
Professor Elizabeth Stasny, Advisor
Professor Catherine Calder
Professor Eloise Kaizar
Professor Asuman Turkmen
Crime rates for small geographic areas, or domains, are often of interest in research applications. However, survey data on victimization is often not reliable at the local level; the main source of crime data in the United States, the National Crime Victimization Survey (NCVS) is designed to produce national estimates only and the public-use data does not contain county identifiers. Even with county identifiers, crime is such a rare event that most sampled counties contain only a handful of victimizations at best. Crime data collected from police reports in the United States, such as in the Uniform Crime Reports (UCR) and the National Incident-Based Reporting System (NIBRS), is widely available at the county level but excludes crimes not reported to the police. The UCR and NIBRS are meant to be censuses of all police-reported crime, but are plagued by missing and incomplete data that makes estimation challenging.

This dissertation presents two methods for estimating county-level crime rates that account for both crimes reported and not reported to the police. In both methods, we take police-reported crime data from NIBRS to calculate county-level crime rates including only crimes reported to the police, then use NCVS data to estimate the percentage of crimes reported to the police in each county of interest. The difference between the two
methods is the mechanism for matching NCVS records to counties, since there is no natural way of linking the two.

In the first method, a resampling technique is used on NCVS records. We use American Community Survey (ACS) five-year estimates to determine the population of each county in sex by race categories, then sample NCVS records according to these proportions. We then use the sampled records to estimate the county-level police reporting rate. A negative binomial distribution can then be used to model the true number of crimes committed in a county, taking the estimated police reporting rate as the probability of success and the number of police-reported crimes as the number of observed successes.

The second method is an adaptation of a hierarchical Bayes model with county-level priors based on the demographic profile of each county to estimate county-level police reporting rates, updated with estimates based on the NCVS geographic identifiers available for each county. The method again uses a negative binomial to model the distribution of the number of all crimes committed in a county. Both methods are illustrated using the crime of aggravated assault, but can be extended to other crime types. These methods can be further extended to other scenarios when small area estimates are desired, if high-quality survey data cannot be used alone because of its limited coverage or sparsity and a large administrative dataset is available but may be biased or covers only a portion of the population of interest.
Dedication

To my parents and to my husband—you made all this possible.
Acknowledgments

First, to my advisor, Dr. Elizabeth A. Stasny: thank you for your guidance, for your critical eye (especially when editing!), and for your enthusiasm and support. I probably would not have made it through two babies to graduation without your flexibility and encouragement. You kept me on track, you set me up with my dream job, and I am so honored to have you as my advisor.

Also to all the professors I’ve had over these five years, particularly my committee members who stepped in at the last minute to help me graduate: thank you. I can’t imagine that any other university has a group of faculty so genuinely committed to the success of their graduate students.

Finally, thank you to my parents for supporting me every step of the way, for believing in me without pressuring me, and for watching the kids so I could work uninterrupted. On our next visit, I won’t have to write anymore! And to my husband and my three children—it’s been a hard road, a crazy road, but we did it. I couldn’t be more excited to set out on our next adventure together.
Vita

June 2003 .................................................. Westerville North High School

May 2007 ................................................. B.A. Mathematics, University of Notre Dame

2012 ....................................................... M.S. Statistics, The Ohio State University

2011 to 2013 ......................................... Graduate Teaching Assistant, Department of Statistics, The Ohio State University

2013 to 2014 ......................................... Graduate Research Assistant, Statistical Consulting Service, The Ohio State University

2015 to present ...................................... Graduate Fellow, The Ohio State University

Fields of Study

Major Field: Statistics
Table of Contents

Abstract ........................................................................................................................................... ii

Dedication ........................................................................................................................................ iv

Acknowledgments ....................................................................................................................... v

Vita................................................................................................................................................ vi

Table of Contents ....................................................................................................................... vii

List of Tables ................................................................................................................................ xi

List of Figures ............................................................................................................................. xiv

Chapter 1: Introduction ............................................................................................................... 1

Chapter 2: Data ............................................................................................................................ 7

   Section 2.1: The National Crime Victimization Survey ......................................................... 7

      History of the NCVS ............................................................................................................... 7

      NCVS Sampling Methodology .......................................................................................... 10

      NCVS Survey Instruments ............................................................................................... 16

      Creation of NCVS data-year file and relevant descriptive data analysis ...................... 19

   Section 2.2: The National Incident-Based Reporting System ........................................... 24
Section 5.5: Hierarchical Bayes-based procedures and results ......................... 162

Method A: Single variance parameter ............................................................. 162

Method B: One variance parameter per MSA by region cell ......................... 171

Method C: One variance parameter per MSA status .................................. 175

Methods D and E: One variance parameter per region ............................... 177

Method F: Two variance parameters assigned by region ............................ 180

Chapter 6: Conclusions and directions for future research .......................... 186

References .................................................................................................... 194

Appendix A: Derivation of appropriate negative binomial distribution ........... 200

Appendix B: Map of Ohio counties, labeled with county names .................... 203

Appendix C: WinBUGS model code ............................................................... 204

Method A ...................................................................................................... 204

Method B ...................................................................................................... 205

Method C ...................................................................................................... 205

Models D, E, and F ...................................................................................... 206

Appendix D: Selected diagnostic plots from Chapter 5 model output .............. 207

Method A ...................................................................................................... 207

Method B ...................................................................................................... 209

Method C ...................................................................................................... 212
Method D ......................................................................................................................... 216
Method E ......................................................................................................................... 221
Method F ......................................................................................................................... 226
List of Tables

Table 1: Example NCVS panel rotation chart ................................................................. 14
Table 2: Month of Interview by Month of Reference (X's denote months in the 6-month
reference period) ........................................................................................................... 15
Table 3: NCVS personal victimizations in data year 2011 by type of crime code .......... 23
Table 4: Incidents, Offenses, and Victims by Offense Category, all NIBRS 2011 data .. 28
Table 5: 25 largest NIBRS agencies by population, as of 2011 ................................. 32
Table 6: Number of victims by type of offense in final NIBRS 2011 analysis file ....... 35
Table 7: 2011 agency-level NIBRS coverage, by region, MSA status, and population
grouping ......................................................................................................................... 38
Table 8: Count of aggravated assaults in the 2011 NCVS by age, sex, and race of victim
........................................................................................................................................ 48
Table 9: Results from weighted logistic regression predicting aggravated assault
victimization in the 2011 NCVS ..................................................................................... 78
Table 10: NCVS 2011 aggravated assaults, by large area category ............................. 81
Table 11: Parameter estimates for Zero-Inflated Beta regression, all counties and
Midwest only .................................................................................................................. 87
Table 12: Parameter estimates from linear model with response on logit scale,
Midwestern counties only ............................................................................................. 93
Table 13: Distribution of NCVS interviews by sex and race of respondent, for all respondents and respondents reporting an aggravated assault only, unweighted. 107
Table 14: Available NCVS Interviews and Interviews with Aggravated Assaults. 113
Table 15: Distribution of NIBRS-only and estimated aggravated assault rates per 1000 people, reporting rates in percentages. Ohio only (n=86 counties). 122
Table 16: Distribution of NIBRS-only and estimated aggravated assault rates per 1000 people, reporting rates in percentages. Midwest only (n=551 counties) 127
Table 17: Distribution of NIBRS-only and estimated aggravated assault rates per 1000 people, reporting rates in percentages. All available U.S. counties (n=1438) 128
Table 18: NCVS 2011 aggravated assaults, by large area category 133
Table 19: Pearson correlations between police reporting rate for aggravated assault and police reporting rate for selected crimes, overall and by MSA and by region. Bold text indicates the correlation is significant at the 95% level. 137
Table 20: NCVS 2011 burglaries by large area category 139
Table 21: Fay-Herriot model comparison 144
Table 22: Distribution of NIBRS-only, simulation-based estimates, and negative binomial model-based estimates of aggravated assault rates per 1000 people and police reporting rates in percentages. Ohio only (n=86 counties) 148
Table 23: Distribution of NIBRS-only, simulation-based estimates, and negative binomial model-based estimates of aggravated assault rates per 1000 people and police reporting rates in percentages. Midwest only (n=555) 150
Table 24: Distribution of NIBRS-only, simulation-based estimates, and negative binomial model-based estimates of aggravated assault rates per 1000 people and police reporting rates in percentages. All available U.S. counties (n=1446) ........................................ 150

Table 25: Aggravated assault rates per 1000 people, reporting rates in percentages. Ohio only (n=86). ................................................................. 160

Table 26: Aggravated assault rates per 1000 people. Ohio only (n=86) .................. 170

Table 27: Aggravated assault police reporting rates in percentages. Ohio only (n=86) 171

Table 28: Count of number of counties per cell, with index and MSA/region labels .... 173

Table 29: Summary statistics for number of NIBRS aggravated assaults per county by MSA status (n=1446) ............................................................................ 177

Table 30: Comparison of all six methods from Section 5.5 ................................. 182
List of Figures

Figure 1: Percentage of population covered by NIBRS, by county, 2011 .................. 30

Figure 2: Distribution of ACS population percentages for counties with NIBRS data
(n=1452) .................................................................................................................. 43

Figure 3: Aggravated assault police reporting rate for geographic cells, 2011 NCVS.... 82

Figure 4: Aggravated assault police reporting rate, adjusted and unadjusted, 2011 NCVS
.................................................................................................................................. 84

Figure 5: Map of Ohio county-level crime rates, from NIBRS and from zero-inflated beta
model......................................................................................................................... 89

Figure 6: Histogram of logit of adjusted NIBRS rates ............................................ 92

Figure 7: Diagnostic plots from linear model, Midwest counties only...................... 94

Figure 8: Map of Ohio county-level crime rates, from NIBRS and using Model 2 ....... 95

Figure 9: Histograms of Simulated Police-Reporting Rates, Unweighted, Run 1 Only vs.
Five Run Average .................................................................................................... 116

Figure 10: Estimated Police-Reporting Rates for Central City and non-Central City
counties, Weighted vs. Unweighted.......................................................................... 118

Figure 11: Maps of Ohio Counties showing NIBRS Aggravated Assault Rates, Estimated
Actual Aggravated Assault Rates, and Estimated Police-Reporting Rates .............. 124

Figure 12: Police reporting rate of selected crime types, NCVS 1993-2013............. 136
Figure 13: Police reporting rate of selected crime types by region and MSA status, NCVS 1993-2013 .......................................................... 138

Figure 14: Fay-Herriot model comparison .......................................................... 143

Figure 15: Histograms of prior and posterior county-level means and variances for the
distribution of reporting rate .......................................................... 156

Figure 16: Sample of 100 beta distributions (out of 1,446 total) using the estimated
parameters .......................................................... 157

Figure 17: Plot of beta posterior distributions of county-level police reporting rate
(n=1,446) .......................................................... 158

Figure 18: Diagnostic plots for checking convergence of precision parameter \( \tau \) (tau.ncvs),
Method A .......................................................... 166

Figure 19: Summary diagnostic plot for checking convergence of selected county-level
means (logit scale), Method A .......................................................... 167

Figure 20: Histograms of county-level means and variances for prior distribution of
\( \rho_j(i) \), posterior estimates for \( \rho_j(i) \) from beta-negative binomial model (Section 5.4), and
posterior estimates for \( \rho_j(i) \) from hierarchical Bayes-based model (this section) ........ 169

Figure 21: Diagnostic plots for checking convergence of \( \tau \) (tau.ncvs), Method B ....... 173

Figure 22: Diagnostic plots for precision parameter for suburban Northeast, Method B174

Figure 23: Diagnostic plots for checking convergence of precision parameter (tau.ncvs),
Method C .......................................................... 176

Figure 24: Diagnostic plots for checking convergence of precision parameter (tau.ncvs),
Method D .......................................................... 178
Figure 25: Counties with fewer than 10 NIBRS aggravated assaults, by region........ 179
Figure 26: Diagnostic plots for checking convergence of precision parameter (tau.ncvs), Method F .............................................................................................................. 180
Figure 27: Aggravated assault rates for Ohio counties, large area model, beta prior model, and Method F .............................................................................................................. 184
Figure 28: Diagnostic plot for deviance, Method A ...................................................... 207
Figure 29: Diagnostic plot for posterior for Franklin County, OH under Method A .... 208
Figure 30: Summary diagnostic plot for selected county-level means, Method B .... 209
Figure 31: Diagnostic plot for deviance, Method B ...................................................... 209
Figure 32: Diagnostic plot for tau.ncvs (variance parameter by large area), Method B. 210
Figure 33: Diagnostic plot for selected tau.ncvs (variance parameter for rural Midwest), Method B .............................................................................................................. 210
Figure 34: Diagnostic plots for posterior for Franklin County, OH under Method B .... 211
Figure 35: Summary diagnostic plot for selected county-level means, Method C .... 212
Figure 36: Diagnostic plot for deviance, Method C ...................................................... 212
Figure 37: Diagnostic plot for tau.ncvs by MSA status, Method C............................... 213
Figure 38: Diagnostic plot for tau.ncvs[1] (variance parameter for urban counties), Method C .............................................................................................................. 213
Figure 39: Diagnostic plot for tau.ncvs[2] (variance parameter for suburban counties), Method C .............................................................................................................. 214
Figure 40: Diagnostic plot for tau.ncvs[3] (variance parameter for rural counties), Method C .............................................................................................................. 214
Figure 59: Diagnostic plot for deviance, Method F ..................................................... 226

Figure 60: Diagnostic plot for tau.ncvs[1] (variance parameter for Northeast and Midwest), Method F ...................................................................................................................... 227

Figure 61: Diagnostic plot for tau.ncvs[2] (variance parameter for South and West), Method F ...................................................................................................................... 227

Figure 62: Diagnostic plots for posterior for Franklin County, OH under Method F .... 228
Chapter 1: Introduction

Large-scale sample surveys, especially government surveys like the American Community Survey (ACS) or National Crime Victimization Survey (NCVS), have been faced with declining response rates, rising costs, and budget cuts in recent years. At the same time, researchers are interested in estimation for small domains such as U.S. counties or specific demographic subgroups. Most large sample surveys are designed to produce reliable estimates at the national level, with state-level estimates available in certain cases, but very few surveys have a large enough sample for county-level estimation based on survey data alone. One proposed solution is to augment survey data with administrative data, such as tax records, Medicare data, or court records.

Administrative data is usually relatively cheap to access since it is already being collected for other purposes, but it often lacks the level of detail in survey data. Caution also must be used when combining data from two different sources; naïve methods can lead to estimates that actually have larger errors than estimates using the survey data alone (see Lohr and Brick 2012, Ybarra and Lohr 2008, or Elliott and Davis 2005).

We are interested in improving estimation of crime rates at the county level. Accurate crime rates are important for policymakers, for news organizations, and for people choosing where to live, among others. Crime trends are often more important than crime...
rates from a policy standpoint: is crime going up or down in an area, and why? What policy changes have been or could be enacted to affect crime? We focus on the county level because many laws which are thought to affect crime rates are enacted at the city or county level: concealed carry regulations or changes in drug crime enforcement, for example. By “crime rates,” we mean not only police-reported crime, but also unreported crime. Looking at police blotters alone can be misleading, because low crime rates based on police reports could simply mean that residents in an area hesitate to call the police; a decrease in police-reported crime rates could be entirely due to a change in law enforcement policy. Intuitively, if ten robberies happened in neighborhood A and all of them were reported to police, but in neighborhood B ten robberies occurred and only two were reported to police, most people would consider both neighborhoods to have the same crime rate—but if looking at only reported crime, a researcher would conclude that neighborhood B has far fewer robberies. This makes the problem more complicated because reliable data on unreported crime are not widely available at the county level, but we believe the resulting estimates will be far more useful.

The two most commonly used sources of crime data for the United States are the National Crime Victimization Survey (NCVS) and the Uniform Crime Reports (UCR)/National Incident-Based Reporting System (NIBRS). The NCVS is a large national survey conducted via interviews of all persons 12 years of age and older living in sampled housing units or group quarters and is administered by the U.S. Census Bureau for the Bureau of Justice Statistics (BJS). The sample is a multi-stage, stratified cluster
sample. The NCVS uses a rotating panel design, under which an interview (either via telephone or in person) is attempted with each eligible household member every six months over a period of three and a half years for a total of seven possible interviews; after seven interviews, the housing unit is rotated out of the sample. Respondents are asked to report any incidents which occurred in the last six months, even if the respondent did not necessarily consider it a crime or report the incident to the police. The NCVS provides estimates of crime rates at the national level and for certain states, but is not a reliable data source for inference about local crime trends, such as the effect of gun control legislation enacted at the state or local level.

NIBRS is not a survey, but a collection of administrative records on crime data. Law enforcement agencies that participate in NIBRS record detailed information on each crime, such as the type of crime, victim characteristics like age, sex, and race, and offender information if known. NIBRS can be considered a census of all police-reported crime for agencies that report through NIBRS. Then, by aggregating all NIBRS crimes reported within a county, it is possible to calculate the true number of police-reported crimes for that county (plus or minus some measurement error, but free of sampling error).

Much of the previous work combining NCVS data with police-reported crime data for small-area estimation uses UCR crime rates to predict NCVS crime rates in a linear regression model, including other predictors like poverty rate or proportion renter as a
proxy for rate of unreported crime (for example, Fay and Li 2011; Fay and Diallo 2012; Rosenfeld 2007). This strategy requires access to the NCVS county-level identifiers, which are not publicly available. Any direct modeling can also only use the several hundred counties that have NCVS sample in a given year out of over 3000 total counties in the United States. Since only about half of all counties have usable NIBRS data, even if we could obtain county-level identifiers on the NCVS data we would likely be building models based on a subset of fewer than 500 counties which would almost certainly not be representative of all U.S. counties, especially given that NIBRS does not cover most of the largest U.S. counties. It might be possible to get stratum information from BJS to determine what counties were grouped together in a stratum for selection, but even then it would be necessary to make some strong assumptions about the relationship between crime rates for counties within a stratum. BJS itself is not confident that their current method of dividing counties results in the most homogeneous strata: the purpose of Fay and Li’s research was to improve the efficiency of the NCVS in part by redefining the way strata are chosen. We want to develop a method that will allow us to use the publicly available data from as many counties as possible while still linking the NCVS data to counties in a meaningful way.

Since the number of police-reported crimes at the county-level is known when NIBRS data is available, we only need to estimate crimes not reported to the police at the county-level. We propose two different methods for estimation. In Chapter 4, we employ a simulation procedure using the NCVS data to estimate the percentage of crimes reported
to the police for each county (the county-level police-reporting rate), which is an extension of a technique used by Calder et al. (2008) to estimate mean particulate matter exposure at the county level. We can then estimate the total county-level number of crimes (reported and unreported), and use ACS estimates of county population to calculate an estimated county-level crime rate. Chapter 5 estimates county-level police reporting rate via a method based on a hierarchical Bayes model. We use the demographic characteristics of the county to construct a prior distribution for county-level police reporting rate based on NCVS data, then update this prior with geographically-based NCVS data. We can again use the county-level estimated police reporting rate along with the NIBRS reported crime rate to estimate the total county-level crime rate. Both methods take advantage of the strengths of each data source. The main advantage of NIBRS is that it is easily available at the county level, but the drawback is that it only covers reported crimes. The NCVS provides data on reported and unreported crimes, but is not reliable at the county level. Rather than attempting to reduce NCVS data to the county level, we use NIBRS data at the county level and augment it with NCVS data from a higher level.

Chapter 2 provides descriptions of each data source used in this research. The NCVS and NIBRS sections include a comprehensive overview of the history and methodology of each data source, as well as detailed instructions on how the analysis files for each data source were constructed and descriptive data analysis for each file. The American Community Survey (ACS) section is less detailed since ACS data is only used briefly in
this research, but contains a summary of ACS methodology and a description of the data files used. Chapter 3 is a review of previous and related work by other researchers, as well as a review of our early work on this problem. In Chapter 4, we present the simulation-based method for estimating county-level crime rates; in Chapter 5, we present the method based on a hierarchical Bayes model. Finally, Chapter 6 contains concluding remarks and directions for future research.
Chapter 2: Data

Section 2.1: The National Crime Victimization Survey

History of the NCVS

The National Crime Victimization Survey (NCVS) is the current version of the former National Crime Survey (NCS), which was launched in 1972 at the recommendation of a special presidential commission. (For more details, see Rennison and Rand's excellent history of the NCVS in the first chapter of Lynch and Addington (2007).) Prior to the introduction of the NCS, crime researchers had to rely exclusively on official police crime data. This posed a problem; not only was it a challenge to estimate the rate of unreported crime, but often variation in police-reported crime rates is more closely tied to shifts in law enforcement policy than to change in actual crime trends. For example, a police commissioner could direct officers to classify all but the most brutal assaults as “simple” rather than “aggravated” assaults. Aggravated assault is considered a violent crime, while simple assault is not—so the city's violent crime rate could appear to drop substantially although the actual amount of violence taking place within the city did not change.

The first NCS was composed of four separate surveys; of these, only the national household survey known as the Crime Panel survived after 1976 and became
synonymous with the NCS name. (The other three surveys were a Central City survey of households, and national and Central City surveys of businesses. The Central City oversamples did not add enough information to justify the expense, and the business surveys did not find much unreported crime—police reports typically cover most crimes committed against businesses.) The Law Enforcement Assistance Administration (LEAA) was formed as a special agency to establish the NCS, and LEAA in turn commissioned the Census Bureau for the actual survey design and implementation.

The NCS underwent several minor (“non-rate-affecting”) changes over the next decades. Most notably, in the late 1970's, the responsibility for the NCS was shifted from LEAA to the Bureau of Justice Statistics (BJS). In 1992, the NCS underwent its first major redesign. The crime screener portion of the survey, designed to identify potential victimizations, was entirely redone to increase reporting of sensitive crimes like rape and sexual assault, as well as to better capture minor crimes like simple assaults or petty theft, all traditionally underreported in the NCS especially when committed by close friends or family members. The pace of the interview was slowed down to give respondents more opportunities to recall victimizations, while prompts were reworded to contain more contextual cues and avoid police terms like assault or larceny. Domestic violence was specifically addressed, and more emphasis was put on screening for crimes committed by people known to the victim. A typical question from the redesigned screener is below; note the lack of legal terms, and the specific mention of incidents committed by friends or relatives.
42a. People often don’t think of incidents committed by someone they know. (Other than any incidents already mentioned,) did you have something stolen from you OR were you attacked or threatened by –

(a) Someone at work or school -

(b) A neighbor or friend -

(c) A relative or family member -

(d) Any other person you've met or known? (NCVS 2011).

Theft, which had been classified as “personal” or “household” (and thus given either a personal-level or household-level survey weight) depending on what property was stolen, was now classified as a purely household crime unless personal contact was involved, as in pocket picking or purse snatching. Computer-assisted telephone interviewing (CATI) was phased in for 30% of interviews each month. CATI was thought to reduce interviewer error, because it forced interviewers to read each question and automated the often complicated skip logic pattern. With CATI, it is also possible to record the amount of time an interviewer spends talking to each respondent and flag interviewers with suspiciously short interview times, which tends to deter interviewers from rushing through interviews and therefore is thought to increase the number of crimes reported.

Because all these changes tended to increase the number of victimizations reported and created a major break in series from previous NCS data, the redesigned survey was also renamed the National Crime Victimization Survey. This change emphasized both the lack of comparability between pre- and post-redesign estimates and the increased focus on
capturing all victimizations, even incidents that victims didn't necessarily perceive as a crime. Emphasizing victimization was a political asset, too: at the time, President Reagan was trying to cut government spending, and survey budgets were easy to cut because most surveys were widely perceived as big government invading citizens’ private lives (E. Stasny, personal communication, July 15, 2014). Rebranding itself as a survey that helped innocent crime victims, rather than a survey about crime or criminals, made it easier for BJS to justify continued funding for the NCVS (although the NCVS still suffered several rounds of budget cuts and sample reductions over the following years).

**NCVS Sampling Methodology**

The current NCVS sampling methodology is largely unchanged from the original NCS methodology, apart from variations in sample size over the years. The NCVS is conducted via personal interviews collected by sampling housing units and group quarters from across the United States. Armed Forces members in military barracks, the homeless, and those in institutions (such as prisons or mental hospitals) are specifically excluded, but persons living in dormitories and religious dwellings are included in the sampling frame (BJS, 2008). All persons ages 12 and older in a sampled housing unit are eligible to be interviewed. In 2011, BJS reported that 79,800 households were interviewed, comprising about 143,120 personal interviews. The response rates for the NCVS are historically quite high—in 2011, the household response rate was 90%, and

---

1 This is a huge increase from 2006, when the sample size dropped down to 38,000 households until a major sample reinstatement in 2010.
the person-level response rate, calculated within participating households only, was 88% (Truman and Planty, 2011).

The NCVS uses a stratified, multistage cluster design. In the first stage, the United States is divided into primary sampling units (PSUs) comprising counties, groups of counties, or large metropolitan areas. The largest PSUs (called “self-representing” PSUs) are included in the sample automatically, each forming its own stratum. The remaining PSUs (“non-self-representing”) are divided into strata based on characteristics including region, population density, and population growth rate as measured at the most recent decennial census, and one PSU is selected from each stratum. These PSUs and strata divisions are revised after each census; the latest revision based on the 2000 census was phased in starting in 2005. The 1990 design included 93 self-representing PSUs and 152 non-self-representing PSUs, which was later reduced to the same 93 self-representing PSUs but only 110 non-self-representing PSUs following a sample reduction in October of 1996 (BJS 2006, 2008). Updated numbers for the 2000 design were not publicly available as of this writing; even the most recent literature published on NCVS methodology refers to the 1990 design. (See, for example, Chapter 7 in Kruttschnitt et al. 2014.)

Next, each selected PSU is divided into four frames which are intended to be non-overlapping. The unit-level frame and the eligible group quarters frame are drawn from the most recent Census address files. However, since new housing units may be constructed after the Census file was generated, the NCVS also draws from an area-level
frame and a permit frame. The area frame contains census blocks based on the most recent census, and the permit frame is drawn from a list of all new building permits. (It is not entirely clear from the publicly available documentation how BJS ensures that a housing unit included in the permit frame is not also included in the area frame.) In each selected PSU, clusters of approximately four housing units are selected from each frame. These selected addresses are listed and given to interviewers for contact (BJS, 2008).

Respondents are interviewed directly about any potential victimizations over the past six months, with three exceptions: 12- or 13-year old persons when a parent or guardian objects to a direct interview, persons who are incapacitated, or persons who are away from the household (such as away at school or on a trip) for the entire field interview period. In those cases, a proxy interview is administered in which a knowledgeable adult responds on behalf of the person. First interviews are traditionally conducted in person, with nearly all follow-up interviews conducted via phone as a cost-saving measure; recently, due to persistent budget cuts even first interviews are often conducted by telephone. In-person interviews were completed using paper-and-pencil questionnaires exclusively until 2006, when computer-assisted personal interviewing (CAPI) was introduced\(^2\). Beginning in 2007, phone interviews were also conducted with interviewers using CAPI to record responses (BJS, 2008).

\(^2\) This sudden change, along with several other changes which were not intended to be rate-affecting but ended up significantly impacting crime rate estimates, caused BJS to declare a break in series for the 2006 NCVS estimates. Many of the changes were reversed in 2007 (although use of CAPI was not), so NCVS estimates from 2007 onward are consistent with estimates from 2005 and earlier, but crime rates from the 2006 NCVS are not comparable to other NCVS estimates and should be used with extreme caution.
Once sampled, a housing unit remains in the sample for seven interviews: one every six months over a period of three years. The first interview was originally used only as a bounding interview to ensure that households were not reporting crimes that occurred before the six-month reference period and were not included in estimates; by 2007, the NCVS sample had been cut so drastically that BJS decided to begin using the first interview data as well (BJS, 2008). Note that the housing unit, not the household, is sampled. This means that if a household moves out in the middle of the sample period and a new household moves in, the new household would be interviewed at all subsequent time points. Alternately, the makeup of the household could change: household members could divorce or marry, or adult children or parents could move in or out of the home, among other possible changes. In all these cases, the relevant changes would be recorded on the household control card, but interviewers would continue to visit the housing unit every six months until seven total interviews were completed for the housing unit.

The NCVS uses a rotating panel design which staggers interviews throughout the year. The sample is divided into six rotation groups, and each rotation group is further divided into six panels. A rotation group will all be on the same interview number, but each panel within the rotation group will be interviewed on a different month. For example, a rotation group could be on its second interview; panel 1 will be interviewed in January, panel 2 will be interviewed in February… panel 6 will be interviewed in June, and in July...
panel 1 will begin the third interview. The rotation groups are staggered so that in any
given month, one-sixth of the sample is on the first interview, one-sixth on the second
interview, and so on. Table 1 is a diagram of what an NCVS rotation panel chart might
look like for a single calendar year. Let the letters (A, B, C…) denote the panels, and the
numbers 1-6 denote the rotation groups within a panel, so that “B3” would denote panel
B, rotation group 3. Notice that the households in panel F complete their 6th interview in
the first half of the year, so in the second half of the year they are rotated out of the
sample and replaced with a new sample of households, panel A*.

<table>
<thead>
<tr>
<th>Interview Month</th>
<th>Interview Number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>January</td>
<td>A1</td>
</tr>
<tr>
<td>February</td>
<td>A2</td>
</tr>
<tr>
<td>March</td>
<td>A3</td>
</tr>
<tr>
<td>April</td>
<td>A4</td>
</tr>
<tr>
<td>May</td>
<td>A5</td>
</tr>
<tr>
<td>June</td>
<td>A6</td>
</tr>
<tr>
<td>July</td>
<td>A1</td>
</tr>
<tr>
<td>August</td>
<td>A2</td>
</tr>
<tr>
<td>September</td>
<td>A3</td>
</tr>
<tr>
<td>October</td>
<td>A4</td>
</tr>
<tr>
<td>November</td>
<td>A5</td>
</tr>
<tr>
<td>December</td>
<td>A6</td>
</tr>
</tbody>
</table>

Table 1: Example NCVS panel rotation chart

Table 2 below shows for persons interviewed in each month the reference period within a
calendar year, which extends six months before the interview month. A respondent
interviewed in January of 2015 would be asked about victimizations occurring in July through December 2014; someone interviewed in June of 2015 would be asked about the period from December 2014 through May 2015.

<table>
<thead>
<tr>
<th>Interview Month</th>
<th>Reference period within calendar year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jan</td>
</tr>
<tr>
<td>January</td>
<td></td>
</tr>
<tr>
<td>February</td>
<td></td>
</tr>
<tr>
<td>March</td>
<td>X</td>
</tr>
<tr>
<td>April</td>
<td>X</td>
</tr>
<tr>
<td>May</td>
<td>X</td>
</tr>
<tr>
<td>June</td>
<td>X</td>
</tr>
<tr>
<td>July</td>
<td>X</td>
</tr>
<tr>
<td>August</td>
<td>X</td>
</tr>
<tr>
<td>September</td>
<td>X</td>
</tr>
<tr>
<td>October</td>
<td>X</td>
</tr>
<tr>
<td>November</td>
<td>X</td>
</tr>
<tr>
<td>December</td>
<td>X</td>
</tr>
<tr>
<td>January</td>
<td>X</td>
</tr>
<tr>
<td>February</td>
<td>X</td>
</tr>
<tr>
<td>March</td>
<td>X</td>
</tr>
<tr>
<td>April</td>
<td>X</td>
</tr>
<tr>
<td>May</td>
<td>X</td>
</tr>
<tr>
<td>June</td>
<td>X</td>
</tr>
<tr>
<td>July</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Month of Interview by Month of Reference (X's denote months in the 6-month reference period)
**NCVS Survey Instruments**

The NCVS consists of two main parts: the NCVS crime screener and the incident report. The screener is designed to help the respondent recall any potential victimizations over the past six months. The interviewer prompts instruct the respondent to mention anything that may qualify, even if it may not be a crime. If the respondent indicates that a crime may have occurred, the interviewer fills out a detailed incident report that collects information on where and when the crime occurred, who committed it, whether the crime was reported to the police, and the exact details of the crime. Since the focus of the NCVS is on victimization, the intent is to record the victim’s perception of the incident; no effort is made to verify incident details. BJS staff later use the detailed incident report to classify any victimization(s) under one of 34 detailed “type of crime” (TOC) codes, or to determine that no crime occurred. Attempted crimes and completed crimes are identified separately, but they are combined in all final BJS reports so that, for example, the reported sexual assault rate takes into account both attempted and completed sexual assaults.

The NCVS divides crimes into household or personal crimes. Household crimes include most property crimes—burglary, larceny, and motor vehicle theft—since it can be difficult to determine exactly which household member owns stolen property. Only one person in the household (the reference person) is interviewed about household crimes. Each person in the household is interviewed about personal victimizations, which include all violent crime, simple assault, and theft involving personal contact like pocket picking.
or purse snatching. Because the NCVS is based on interviews with victims, murder and kidnapping are outside the scope of the NCVS.

The NCVS data files contain three types of weights: victimization weights, incident weights, and person/household weights. Nonzero victimization weights are included on each record reporting a victimization, and the appropriate victimization weight is attached to the household level record for household crimes and the person level record for personal crimes. Victimization weights are used to calculate the numerator of all victimization rates reported by BJS, which are what are typically considered crime rates. Incident weights are similar to victimization weights, except that they are designed to calculate the crime incident rate rather than the crime rate; e.g., if two NCVS respondents were robbed at the same time, the victimization weight would be twice as large as the incident weight, because there were two victims but only one incident. Household and person weights are used to calculate denominators for the respective crime rates (BJS 2013). For example, to calculate the aggravated assault victimization rate, one would use the person-level records since aggravated assault is a personal crime, take the sum of all aggravated assaults weighted by the personal victimization weight, and then divide that quantity by the sum of all person weights in the file.

Each weight is composed of five components: a base weight, special adjustments, a household noninterview factor, a first-stage ratio, and a second-stage ratio. Base weights are set with each NCVS sample redesign to be self-weighting based on the
noninstitutionalized population of the United States ages 12 and older; the last redesign
was based on the 2000 Census. The special adjustments account for housing units that are
larger than expected and must be subsampled, such as an apartment building expected to
have two units but the interviewer finds that there are actually four units. In this case, the
interviewer will randomly select two units to sample, and those two units will receive a
special weighting factor of 2 since they also represent the two units not sampled. The
first-stage ratio adjustment is done at the PSU level: all non-self-representing PSUs
within a state are aggregated, and weights are adjusted so that the distribution of
black/non-black respondents is identical to the statewide census estimate of black/non-
black population. (Self-representing PSUs are not included since they only represent
themselves.) In the second-stage ratio adjustment, the weights are adjusted to match
national monthly Census Bureau projections for selected age/sex/race and
age/sex/ethnicity cells (BJS 2014).

All person-level weights also contain a within-household noninterview factor to account
for persons in a selected household who could not be interviewed during the field period.
Victimization and incident weights as of 2007 have a bounding weight applied to the first
interview, because respondents typically report more victimizations at the first interview:
even though the first interview asks respondents to report any victimizations over the past
six months, it is often difficult to recall exactly when a crime happened. Especially with
serious crimes a “telescoping” effect can occur, in which a respondent will think that the
crime happened more recently than it actually did. This is usually not an issue with later
interviews, since the interviewer can simply ask about crimes since the last interview. Finally, personal incident weights include an adjustment for multiple victims: if more than one NCVS respondent reports the same incident, the incident weight is divided among respondents so that it only counts for one incident. For more detail on the NCVS weighting, see the NCVS Technical Documentation (BJS 2014).

Creation of NCVS data-year file and relevant descriptive data analysis

Our research focuses on victimizations occurring in calendar year 2011. Police crime data reported in the Uniform Crime Reports (UCR) and NIBRS are released based on the year in which the crimes occurred—all crimes that occurred between January 1, 2011 and December 31, 2011 are recorded in the 2011 data file. However, BJS no longer releases data-year NCVS files, where a data year is defined as all victimizations reported as occurring in a calendar year. Instead, BJS releases collection-year files that include all victimizations reported in interviews during a calendar year. This means that a person interviewed in January 2011 could be reporting a crime that occurred in July 2010 (within the six-month reporting window for that interview date), but the victimization would count towards the reported 2011 crime rates, not the 2010 crime rates. BJS explains that crime rates calculated on a collection-year basis are consistent with data-year based rates, and using the collection year format allows them to produce reports in a more timely manner. Under the collection year, all data for the 2011 report are collected as of December 31, 2011; if BJS wanted to produce a data year report, all interviews would not
be completed until May 2012 since the six-month recall period for May interviews extends back to the previous December.

The differences in victimization rates between collection-year and data-year formats are minor when looking at crime trends over time, but when comparing NCVS data to police records we would like to look at exactly the same time periods. BJS provides instructions for creating a data-year file in the NCVS codebook along with SPSS code (BJS 2013). This requires downloading both the 2011 and 2012 collection-year files and subsetting them to victimizations which occurred during calendar year 2011. We then adjusted all household and personal weights to account for time in sample during 2011 as described below. In the collection-year file, respondents who are interviewed twice have the opportunity to report twelve months of victimizations, and respondents who are interviewed once can report six months of victimizations. The weights in the collection-year file reflect an adjustment for respondents who were interviewed only once, and no adjustment for respondents interviewed twice. However, under the data-year format, a respondent interviewed for the first time in May 2012 was in the sample for only one month during calendar year 2011; a respondent interviewed for the first time in December 2011 and then for the second time in May 2012 was only in the sample for seven months of 2011, and so on. The NCVS codebook recommends adjusting both the household and personal weights (used to calculate the denominators of crime rates) based on time in sample for the data year.
Finally, an adjustment must be made for a change in how BJS handles series victimizations, a detail that is not noted in the NCVS codebook. An incident is classified as a series victimization when a respondent reports at least six victimizations that are so similar that he or she is unable to recall specific details of each incident. When that occurs, interviewers are instructed to collect detailed information about only the most recent incident, mark the victimization as “series,” and note the number of times the victimization has occurred during the recall period. For example, a woman in a domestic violence situation may be unable to recall details of every time her partner has assaulted her over the past six months.

Traditionally, BJS has simply excluded series victimizations from the NCVS, since series victimizations can seriously impact estimates for rare crimes like rape, and typically have less reliable data: detailed information is not available for each incident, and victims often round to intervals of time such as once a month or once a week. (See Truman 2011 or BJS 2013 for more information on series victimizations in the NCVS.) We noticed a discrepancy between crime rates calculated following the codebook instructions and rates in BJS publications, so we contacted BJS directly for guidance. The published code excludes series victimizations entirely, but beginning in 2011 BJS included series victimizations up to a maximum of 10 victimizations per event in their published rates. Series victimizations that were missing the variable for number of victimizations were assigned a value of 6 victimizations (L. Langton, personal communication, March 12, 2014).
Table 3 shows the number of NCVS personal victimizations reported in data year 2011 by detailed type of crime code and by broad crime type (V4529, “Type of Crime Code - New”). The NCVS data file subdivides victimizations into detailed categories, but in published BJS reports these specific crime codes are nearly always aggregated into more general crime types. Attempted and completed crimes are aggregated in NCVS reports, as a general practice. Victimizations which occurred in the Midwest only are displayed separately as well, since certain analyses in later chapters use only data from the Midwest. Percent reported to the police is included for the full sample only. Serious crimes like robbery and aggravated assault tend to be reported to the police at relatively high rates, while sexual crimes and minor crimes like simple assault have much lower police reporting rates.

The data-year approach results in estimated police-reporting rates that are comparable to collection-year numbers published by BJS, and in most cases nearly equal. The same is true for crime rates. For example, using the calendar-year data above, we estimate a national rate of 0.9 rapes and sexual assaults, 2.4 robberies, 4.1 aggravated assaults, and 16.2 simple assaults per 1000 persons. The corresponding numbers in Criminal Victimization 2011, which publishes crime rates based on the 2011 collection year, are 0.9 rapes and sexual assaults, 2.2 robberies, 4.1 aggravated assaults, and 15.3 simple assaults per 1000 persons. Since these national rates are so close, we use selected BJS subnational estimates as a benchmark for our methods, even though BJS uses collection-year rather than data-year NCVS data.
<table>
<thead>
<tr>
<th>Type of crime-NCVS code¹</th>
<th>General crime type</th>
<th>Unweight-ed, full sample</th>
<th>Weighted, full sample</th>
<th>Percent reported to police, full sample</th>
<th>Unwtd, Midwest only</th>
<th>Weighted, Midwest only</th>
</tr>
</thead>
<tbody>
<tr>
<td>(01) Completed rape</td>
<td>Rape</td>
<td>22</td>
<td>78649.59</td>
<td>33.68%</td>
<td>7</td>
<td>24831.05</td>
</tr>
<tr>
<td>(02) Attempted rape</td>
<td></td>
<td>17</td>
<td>57916.77</td>
<td>47.41%</td>
<td>4</td>
<td>12957.03</td>
</tr>
<tr>
<td>Rape total</td>
<td></td>
<td>39</td>
<td>136566.36</td>
<td>39.51%</td>
<td>11</td>
<td>37788.08</td>
</tr>
<tr>
<td>(03) Sex aslt w s aslt</td>
<td>Sexual assault</td>
<td>5</td>
<td>12282.41</td>
<td>34.96%</td>
<td>3</td>
<td>9177.63</td>
</tr>
<tr>
<td>(04) Sex aslt w m aslt</td>
<td></td>
<td>1</td>
<td>2878.31</td>
<td>100.00%</td>
<td>1</td>
<td>2878.31</td>
</tr>
<tr>
<td>(15) Sex aslt wo inj</td>
<td></td>
<td>11</td>
<td>53993.18</td>
<td>24.08%</td>
<td>1</td>
<td>26807.41</td>
</tr>
<tr>
<td>(16) Unw sex wo force</td>
<td></td>
<td>2</td>
<td>4708.99</td>
<td>0.00%</td>
<td>2</td>
<td>4708.99</td>
</tr>
<tr>
<td>(18) Verbal thr rape</td>
<td></td>
<td>4</td>
<td>16458.17</td>
<td>63.01%</td>
<td>1</td>
<td>3829.86</td>
</tr>
<tr>
<td>(19) Ver thr sex aslt</td>
<td></td>
<td>4</td>
<td>14989.98</td>
<td>26.90%</td>
<td>1</td>
<td>4307.95</td>
</tr>
<tr>
<td>Sexual assault total</td>
<td></td>
<td>27</td>
<td>105311.04</td>
<td>32.83%</td>
<td>9</td>
<td>51710.16</td>
</tr>
<tr>
<td>(05) Rob w inj s aslt</td>
<td>Robbery</td>
<td>15</td>
<td>50078.90</td>
<td>71.92%</td>
<td>2</td>
<td>4759.12</td>
</tr>
<tr>
<td>(06) Rob w inj m aslt</td>
<td></td>
<td>19</td>
<td>61222.82</td>
<td>59.07%</td>
<td>2</td>
<td>3739.57</td>
</tr>
<tr>
<td>(07) Rob wo injury</td>
<td></td>
<td>84</td>
<td>30444.50</td>
<td>68.93%</td>
<td>12</td>
<td>58636.21</td>
</tr>
<tr>
<td>(08) At rob inj s aslt</td>
<td></td>
<td>4</td>
<td>17176.42</td>
<td>74.87%</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>(09) At rob inj m aslt</td>
<td></td>
<td>7</td>
<td>21262.11</td>
<td>83.19%</td>
<td>2</td>
<td>7335.57</td>
</tr>
<tr>
<td>(10) At rob w aslt</td>
<td></td>
<td>35</td>
<td>163240.70</td>
<td>54.92%</td>
<td>7</td>
<td>23024.06</td>
</tr>
<tr>
<td>Robbery total</td>
<td></td>
<td>164</td>
<td>617422.45</td>
<td>65.15%</td>
<td>25</td>
<td>97494.53</td>
</tr>
<tr>
<td>(11) Ag aslt w injury</td>
<td>Aggravated assault</td>
<td>91</td>
<td>373431.10</td>
<td>79.44%</td>
<td>21</td>
<td>104534.80</td>
</tr>
<tr>
<td>(12) At ag aslt w wea</td>
<td></td>
<td>78</td>
<td>332685.40</td>
<td>55.39%</td>
<td>17</td>
<td>48286.92</td>
</tr>
<tr>
<td>(13) Thr aslt w weap</td>
<td></td>
<td>77</td>
<td>344086.80</td>
<td>62.30%</td>
<td>18</td>
<td>74940.69</td>
</tr>
<tr>
<td>Aggravated assault total</td>
<td></td>
<td>246</td>
<td>1050203.30</td>
<td>66.20%</td>
<td>56</td>
<td>227762.41</td>
</tr>
<tr>
<td>(14) Simp aslt w inj</td>
<td>Simple assault</td>
<td>174</td>
<td>747861.30</td>
<td>50.38%</td>
<td>45</td>
<td>224025.70</td>
</tr>
<tr>
<td>(17) Asl wo weap, wo inj</td>
<td></td>
<td>313</td>
<td>1305525.00</td>
<td>39.39%</td>
<td>75</td>
<td>305139.10</td>
</tr>
<tr>
<td>(20) Verbal thr aslt</td>
<td></td>
<td>400</td>
<td>2115695.00</td>
<td>37.33%</td>
<td>103</td>
<td>555635.00</td>
</tr>
<tr>
<td>Simple assault total</td>
<td></td>
<td>887</td>
<td>4169081.30</td>
<td>40.31%</td>
<td>223</td>
<td>1084799.80</td>
</tr>
<tr>
<td>(21) Purse snatching</td>
<td>Personal larceny</td>
<td>13</td>
<td>36766.27</td>
<td>63.35%</td>
<td>2</td>
<td>5352.64</td>
</tr>
<tr>
<td>(22) At purse snatch</td>
<td></td>
<td>3</td>
<td>7992.17</td>
<td>0.00%</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>(23) Pocket picking</td>
<td></td>
<td>42</td>
<td>142627.80</td>
<td>40.44%</td>
<td>9</td>
<td>35235.86</td>
</tr>
<tr>
<td>Personal larceny total</td>
<td></td>
<td>58</td>
<td>187386.24</td>
<td>43.21%</td>
<td>11</td>
<td>40588.50</td>
</tr>
</tbody>
</table>

Table 3: NCVS personal victimizations in data year 2011 by type of crime code

¹ Abbreviations used in Type of Crime codes: Asl, aslt (assault); s aslt, s alt (serious assault); m aslt, m asl (minor assault); at (attempted); rob (robbery); inj (injury); simp (simple); thr (threat); unw (unwanted); weap (weapon); w (with); wo (without).
Section 2.2: The National Incident-Based Reporting System

*NIBRS History and Methodology*

Like the NCVS, the National Incident Based Reporting System (NIBRS) is an updated version of an older system. The Uniform Crime Reports (UCR) began in 1930 with 400 law enforcement agencies sending monthly crime counts to the Federal Bureau of Investigation; today, over 18,000 law enforcement agencies participate in the UCR program (Barnett-Ryan 2007; Maltz 1999). Each agency is asked to report a monthly count of seven “Index Crimes”: murder/manslaughter, forcible rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft. The first four crimes are considered personal crimes, and the last three crime types are property crimes. These crime types were chosen in 1930 because they were considered serious enough that they would be widely recognized as crimes and therefore reported to the police, had an agreed upon definition across law enforcement agencies, and happened frequently enough that meaningful data could be collected (Maltz 1999)\(^3\).

The FBI uses the UCR as its primary source of crime data, publishing the yearly report “Crime in the United States” for national and state level crime trends based on aggregated UCR reports. However, since the UCR program began collecting data, crime researchers have become interested in more fine-grained analysis of crime rates: What do crime

---

\(^3\) There is good evidence that all these points are not actually true—rape, for example, is severely underreported, and its definition varies between jurisdictions and has changed over time. The UCR actually stopped accepting rape counts from the state of Illinois for a period in the 1980s and 1990s because Illinois considered the rape of males a crime, while the FBI definition of rape was limited to “the carnal knowledge of a woman against her will” only (Maltz 1999).
trends look like at the county or local level? What are the characteristics of victims and offenders? The UCR only reports monthly crime counts at the law enforcement agency level; we can find out how many aggravated assaults occurred under the jurisdiction of the Columbus police for every month of 2011, but the UCR cannot provide any information on the race or gender of victims, or even how many victims were involved in each incident. Put another way, UCR data are available only at the crime incident level, defined as “one or more offenses committed by the same offender, or group of offenders acting in concert, at the same time and place” (NIBRS 2011 publication resources). It is not possible to use UCR data at the crime-victim level or at the offense level, since a single incident involving multiple offenses or multiple victims is counted in the UCR the same way as a single incident involving only one offense against one victim.

In response to these shortcomings, the FBI began developing the National Incident-Based Reporting System (NIBRS) in the late 1980s. NIBRS technology allowed extracts from police incident reports to be sent directly to the FBI, typically by an automated process as most police agencies switched over to computerized records. The information collected by NIBRS includes “…the nature and types of specific offenses in the incident, characteristics of the victim(s) and offender(s), types and value of property stolen and recovered, and characteristics of persons arrested in connection with a crime incident” (NIBRS 2011). Although law enforcement agencies still provide reports at the incident level, this additional information permits analysis at the victim or offense level as well. In addition, the public use NIBRS files include detailed information on the reporting
agencies, including what county or counties contain the agency, the total population covered by the agency, and the proportion of the agency's population in each county.

NIBRS' data collection methods also remedy some legacy UCR data collection rules that were intended to ease the administrative burden on police agencies when all UCR data had to be recorded by hand. The Hierarchy Rule instructed agencies to only report the most serious crime that occurred, if several offenses took place in one incident. For example, if a victim was badly beaten and then carjacked, under the Hierarchy Rule only the aggravated assault would be reported since motor vehicle theft is considered a lesser crime\(^\text{4}\). This leads to underreporting of lesser crimes in the UCR—while all murders are reported (murder is considered the most serious crime), the less serious the crime, the more likely it is to be affected by the Hierarchy Rule. The Hotel Rule states that if a crime takes place at a hotel, self-storage facility, or other place where the manager is likely to report the crime, it should be counted as one incident rather than separate incidents. For example, if five hotel rooms are burglarized in one night, the UCR will report one burglary instead of five burglaries. Based on the summary UCR data alone, there is no way to tell how many people were victimized in each incident. Using NIBRS, researchers can decide whether or not to apply these rules. NIBRS would report the first example as one incident including two offenses (aggravated assault and motor vehicle theft), and the second example as one robbery with five individual victims.

\(^{4}\) However, the Separation of Time and Place rule says that if the victim was beaten in one location, forced to drive around for a period of time, and then the car was stolen in a separate location, both the aggravated assault and the motor vehicle theft would be reported to the UCR as two separate incidents.
Table 4 provides the total number of incidents, offenses, and victims by offense category in the 2011 NIBRS public use data file. Notice that the number of offenses is the same as the number of victims for crimes against persons: a person who assaults two victims at the same time will count as one crime incident, but will be charged with two separate assault offenses, one for each victim. Similarly, for crimes against property and crimes against society the number of incidents and the number of offenses are the same, but there could be multiple victims involved in each incident/offense. (See Table 4 for a listing of which crime types fall under each category.)

NIBRS seems to provide a huge amount of crime data, covering nearly 5 million crime incidents in 2011. However, the major limitation of NIBRS is its lack of implementation. Those 5 million incidents are only a fraction of the total number of incidents reported to the police in 2011 because many law enforcement agencies choose not to send reports to NIBRS. As of June 2012, the Justice Research and Statistics Association (JRSA) Incident-Based Reporting Resource Center reported that only 32 states have been NIBRS certified. “NIBRS certified” means only that agencies within the state can choose to submit NIBRS data, and does not indicate that agencies are actually reporting through NIBRS. Alabama is NIBRS certified, but JRSA reports that only 1 agency covering 2% of state crime and 1% of state population is reporting through NIBRS. 43% of law enforcement agencies use NIBRS, but these agencies cover only about 27% of the U.S. population and 29% of police-reported crimes.
<table>
<thead>
<tr>
<th>Offense Category¹</th>
<th>Incidents²</th>
<th>Offenses</th>
<th>Victims</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>4,926,829</td>
<td>5,643,241</td>
<td>5,946,990</td>
</tr>
<tr>
<td>Crimes Against Persons</td>
<td>1,149,923</td>
<td>1,321,523</td>
<td>1,321,523</td>
</tr>
<tr>
<td>Assault Offenses³</td>
<td>1,062,148</td>
<td>1,226,300</td>
<td>1,226,300</td>
</tr>
<tr>
<td>Aggravated Assault</td>
<td>163,156</td>
<td>197,986</td>
<td>197,986</td>
</tr>
<tr>
<td>Homicide Offenses</td>
<td>3,546</td>
<td>3,792</td>
<td>3,792</td>
</tr>
<tr>
<td>Kidnapping/Abduction</td>
<td>13,470</td>
<td>15,310</td>
<td>15,310</td>
</tr>
<tr>
<td>Sex Offenses, Forcible</td>
<td>64,382</td>
<td>69,356</td>
<td>69,356</td>
</tr>
<tr>
<td>Sex Offenses, Nonforcible</td>
<td>6,377</td>
<td>6,765</td>
<td>6,765</td>
</tr>
<tr>
<td>Crimes Against Property</td>
<td>3,687,952</td>
<td>3,687,952</td>
<td>3,991,316</td>
</tr>
<tr>
<td>Arson</td>
<td>15,467</td>
<td>15,467</td>
<td>17,790</td>
</tr>
<tr>
<td>Bribery</td>
<td>293</td>
<td>293</td>
<td>317</td>
</tr>
<tr>
<td>Burglary/Breaking &amp; Entering</td>
<td>575,394</td>
<td>575,394</td>
<td>647,672</td>
</tr>
<tr>
<td>Counterfeiting/Forgery</td>
<td>74,131</td>
<td>74,131</td>
<td>81,407</td>
</tr>
<tr>
<td>Destruction/Damage/Vandalism</td>
<td>810,046</td>
<td>810,046</td>
<td>871,667</td>
</tr>
<tr>
<td>Embezzlement</td>
<td>17,000</td>
<td>17,000</td>
<td>17,635</td>
</tr>
<tr>
<td>Extortion/Blackmail</td>
<td>1,217</td>
<td>1,217</td>
<td>1,329</td>
</tr>
<tr>
<td>Fraud Offenses</td>
<td>245,301</td>
<td>245,301</td>
<td>271,205</td>
</tr>
<tr>
<td>Larceny/Theft Offenses</td>
<td>1,683,877</td>
<td>1,683,877</td>
<td>1,782,977</td>
</tr>
<tr>
<td>Motor Vehicle Theft</td>
<td>161,169</td>
<td>161,169</td>
<td>166,660</td>
</tr>
<tr>
<td>Robbery</td>
<td>72,143</td>
<td>72,143</td>
<td>95,737</td>
</tr>
<tr>
<td>Stolen Property Offenses</td>
<td>31,914</td>
<td>31,914</td>
<td>36,920</td>
</tr>
<tr>
<td>Crimes Against Society</td>
<td>633,766</td>
<td>633,766</td>
<td>634,151</td>
</tr>
<tr>
<td>Drug/Narcotic Offenses</td>
<td>550,343</td>
<td>550,343</td>
<td>550,656</td>
</tr>
<tr>
<td>Gambling Offenses</td>
<td>1,238</td>
<td>1,238</td>
<td>1,238</td>
</tr>
<tr>
<td>Pornography/Obscene Material</td>
<td>5,791</td>
<td>5,791</td>
<td>5,795</td>
</tr>
<tr>
<td>Prostitution Offenses</td>
<td>10,213</td>
<td>10,213</td>
<td>10,215</td>
</tr>
<tr>
<td>Weapon Law Violations</td>
<td>66,181</td>
<td>66,181</td>
<td>66,247</td>
</tr>
</tbody>
</table>

Table 4: Incidents, Offenses, and Victims by Offense Category, all NIBRS 2011 data

² The actual number of incidents is 4,926,829. However, the column figures will not add to the total because incidents may include more than one offense type, and each appropriate offense type is counted in this table.
³ Includes aggravated assault, simple assault, and intimidation.
NIBRS tends to be biased towards medium-sized agencies—the costs of switching to NIBRS are typically higher for small agencies that lack resources, and large agencies faced with major changes to their current computer system. The UCR summary system is also plagued by missing data (for an excellent discussion of this, see Maltz 1999 or Addington and Lynch 2007), but for national or state reporting purposes enough agencies report to generate reasonably reliable crime estimates. Over 90% of agencies report at least one month of crime data to the UCR program, and over 90% of the U.S. population is covered by agencies that report to the UCR.

There is a clear dichotomy between states that have adopted NIBRS and states that have not—in Figure 1 below, one can easily identify the 17 states that have fully adopted NIBRS as of 2011, because most counties in those states have 96% or more of their population covered by NIBRS agencies. This map takes into account agencies that cross county lines—that is, if a NIBRS agency is split across two counties, the covered population is also allocated between the two counties. It corresponds very closely to the published numbers of NIBRS coverage (JRSA 2015a.). States which are reported to be nearly 100% covered by JRSA also appear nearly 100% covered in Figure 1, and vice versa.
Figure 1: Percentage of population covered by NIBRS, by county, 2011
Table 5 provides a listing of the largest agencies participating in NIBRS as of 2011; as of this writing, the population numbers of these agencies have changed somewhat, but no other large agencies have adopted NIBRS (JRSA 2015b). As of June 2012, the largest NIBRS agency by population was Fairfax County, Virginia, covering a little over 1 million people; the second largest NIBRS agency is Columbus, OH, covering nearly 780,000 people. Cincinnati and Cleveland also make the list of the top 25 largest NIBRS agencies. (JRSA 2015b). The West and the South census regions have relatively low NIBRS rates for large cities; places like San Francisco, Los Angeles, Atlanta, and New Orleans do not participate in NIBRS. In the Midwest, the large city missing from NIBRS is Chicago, and most of the other large cities participate in NIBRS. In fact, 8 of the 25 largest NIBRS agencies in 2011 are in the Midwest census region (in bold in Table 5).

NIBRS data is richer and more accurate than summary UCR counts, since information is drawn automatically from electronic incident reports, but its spotty coverage limits its usefulness. The FBI and BJS are actively trying to recruit more agencies, especially large agencies, to NIBRS since implementation has largely stalled. BJS is sponsoring the National Crime Statistics Exchange (NCS-X), which provides grants and assistance to help law enforcement agencies switch to NIBRS reporting, as well as other incentives like optimizing agency resources during NIBRS implementation. The planning stages of NCS-X, including outreach to agencies, developing cost and feasibility guidelines, and creating a sampling frame, were completed during 2013 and early 2014. NCS-X was rolled out in summer 2014 in 400 sampled agencies, after which BJS will assess its
impact and consider extending the program to other non-NIBRS agencies. BJS has made it clear with NCS-X that they view NIBRS as the future of police-reported crime statistics, so it makes sense to rely on NIBRS to develop methods with the assumption that NIBRS coverage will continue to improve over the next several years.

<table>
<thead>
<tr>
<th>Number</th>
<th>State</th>
<th>Agency</th>
<th>Population Covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>VA</td>
<td>Fairfax County</td>
<td>1,055,204</td>
</tr>
<tr>
<td>2</td>
<td>OH</td>
<td>Columbus</td>
<td>787,609</td>
</tr>
<tr>
<td>3</td>
<td>TX</td>
<td>Fort Worth</td>
<td>756,803</td>
</tr>
<tr>
<td>4</td>
<td>MI</td>
<td>Detroit</td>
<td>713,239</td>
</tr>
<tr>
<td>5</td>
<td>TN</td>
<td>Memphis</td>
<td>652,725</td>
</tr>
<tr>
<td>6</td>
<td>TN</td>
<td>Nashville</td>
<td>612,789</td>
</tr>
<tr>
<td>7</td>
<td>CO</td>
<td>Denver</td>
<td>610,612</td>
</tr>
<tr>
<td>8</td>
<td>WI</td>
<td>Milwaukee</td>
<td>597,426</td>
</tr>
<tr>
<td>9</td>
<td>CT</td>
<td>Connecticut State Police</td>
<td>539,137</td>
</tr>
<tr>
<td>10</td>
<td>MO</td>
<td>Kansas City</td>
<td>461,458</td>
</tr>
<tr>
<td>11</td>
<td>VA</td>
<td>Virginia Beach</td>
<td>443,226</td>
</tr>
<tr>
<td>12</td>
<td>CO</td>
<td>Colorado Springs</td>
<td>423,680</td>
</tr>
<tr>
<td>13</td>
<td>DE</td>
<td>New Castle County</td>
<td>407,235</td>
</tr>
<tr>
<td>14</td>
<td>VA</td>
<td>Prince William County</td>
<td>398,550</td>
</tr>
<tr>
<td>15</td>
<td>OH</td>
<td>Cleveland</td>
<td>397,106</td>
</tr>
<tr>
<td>16</td>
<td>KS</td>
<td>Wichita</td>
<td>384,796</td>
</tr>
<tr>
<td>17</td>
<td>WA</td>
<td>Pierce</td>
<td>373,553</td>
</tr>
<tr>
<td>18</td>
<td>MI</td>
<td>Oakland</td>
<td>357,881</td>
</tr>
<tr>
<td>19</td>
<td>SC</td>
<td>Greenville</td>
<td>326,146</td>
</tr>
<tr>
<td>20</td>
<td>VA</td>
<td>Chesterfield County</td>
<td>320,014</td>
</tr>
<tr>
<td>21</td>
<td>VA</td>
<td>Henrico County</td>
<td>301,602</td>
</tr>
<tr>
<td>22</td>
<td>KY</td>
<td>Lexington</td>
<td>297,847</td>
</tr>
<tr>
<td>23</td>
<td>OH</td>
<td>Cincinnati</td>
<td>297,160</td>
</tr>
<tr>
<td>24</td>
<td>UT</td>
<td>Salt Lake County</td>
<td>266,224</td>
</tr>
</tbody>
</table>

Table 5: 25 largest NIBRS agencies by population, as of 2011
Creation of NIBRS 2011 analysis file

We use NIBRS data for calendar year 2011, defined as all NIBRS records available for crimes committed between January 1, 2011 and December 31, 2011. The 2011 NIBRS public use data files can be downloaded from the National Archive of Criminal Justice Data (NACJD) hosted by the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan, available online at https://www.icpsr.umich.edu/icpsrweb/NACJD/. NIBRS data files from 1991 onward are also available through NACJD. The 2011 NIBRS data used in this research are taken from ICPSR record number 34585, which contains 13 separate data files aggregated at various levels. NIBRS data can be used at the agency level, victim level, incident level, offender level, or arrestee level, and these 13 files can be combined to create custom data files at the analysis unit of interest.

For our research, the victim level is the most appropriate unit of analysis since the NCVS is also conducted at the victim level. We built a custom victim-level file by merging agency information from the three batch header files onto the victim level file (DS007) so that crimes could be separated into geographic units based on the agency. Next, victimizations against non-individuals or children under 12 were removed so that this file would correspond as closely as possible to the population covered by the National Crime Victimization Survey (NCVS). Finally, we aggregated the data to the county level, allocating crimes for agencies that cross county lines by population when necessary. In some cases, agencies in the NIBRS file did not have county-level information attached,
but it was possible to manually assign a county to the agency—for example, the state of Virginia recognizes 38 independent cities that are geographically located within a county but are considered legally separate from the county. For the purposes of our research, however, it makes sense to include the city crimes in the surrounding county data.

For counties with only partial NIBRS coverage we use the total number of NIBRS reported aggravated assaults divided by the covered population in the county and assume that the rate in the covered portion is the same as in the non-covered portion. We exclude crimes reported through state police, college police forces, and other special agencies like the Forest Service that are not assigned to a specific county or set of counties because we lack the subject area expertise to allocate such crimes appropriately. Maltz (1999) points out that simply allocating state police crimes based on county population is a bad strategy because typically state police have more duties in rural, low-population counties than in urban areas with their own police forces. In the 2011 NIBRS data, there are only 1099 such aggravated assaults out of 186,287 total aggravated assaults against victims 12 and older; other crime types may be more heavily affected by omitting such agencies. This will result in a slight undercount of police-reported aggravated assaults.

The final data file contains 1,452 county-level records with usable NIBRS data out of approximately 3,142 counties and county equivalents in the United States. Any future references to NIBRS data or 2011 NIBRS data refer to this custom data file unless otherwise specified. Table 6 presents the number of victims included in our analysis data.
file (restricted to only individual victims 12 and older, for consistency with the NCVS sample) compared to all victims in the 2011 NIBRS data file by offense category.

<table>
<thead>
<tr>
<th>Offense category</th>
<th>Number of individual victims 12 and older only</th>
<th>Total number of victims</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL</td>
<td>4,246,751</td>
<td>5,643,241</td>
</tr>
<tr>
<td>Crimes against persons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assault Offenses</td>
<td>1,256,833</td>
<td>1,321,523</td>
</tr>
<tr>
<td>Aggravated Assault</td>
<td>1,181,033</td>
<td>1,226,300</td>
</tr>
<tr>
<td>Homicide Offenses</td>
<td>3,792</td>
<td>3,792</td>
</tr>
<tr>
<td>Kidnapping/Abduction</td>
<td>13,456</td>
<td>15,310</td>
</tr>
<tr>
<td>Sex Offenses, Forcible</td>
<td>52,550</td>
<td>69,356</td>
</tr>
<tr>
<td>Sex Offenses, Nonforcible</td>
<td>6,002</td>
<td>6,765</td>
</tr>
<tr>
<td>Crimes Against Property</td>
<td>2,989,918</td>
<td>3,991,316</td>
</tr>
<tr>
<td>Arson</td>
<td>10,533</td>
<td>17,790</td>
</tr>
<tr>
<td>Bribery</td>
<td>213</td>
<td>317</td>
</tr>
<tr>
<td>Burglary/Breaking &amp; Entering</td>
<td>546,177</td>
<td>647,672</td>
</tr>
<tr>
<td>Counterfeiting/Forgery</td>
<td>39,165</td>
<td>81,407</td>
</tr>
<tr>
<td>Destruction/Damage/Vandalism</td>
<td>671,524</td>
<td>871,667</td>
</tr>
<tr>
<td>Embezzlement</td>
<td>3,733</td>
<td>17,635</td>
</tr>
<tr>
<td>Extortion/Blackmail</td>
<td>1,248</td>
<td>1,329</td>
</tr>
<tr>
<td>Fraud Offenses</td>
<td>204,971</td>
<td>271,205</td>
</tr>
<tr>
<td>Larceny/Theft Offenses</td>
<td>1,250,021</td>
<td>1,782,977</td>
</tr>
<tr>
<td>Motor Vehicle Theft</td>
<td>154,725</td>
<td>166,660</td>
</tr>
<tr>
<td>Robbery</td>
<td>82,945</td>
<td>95,737</td>
</tr>
<tr>
<td>Stolen Property Offenses</td>
<td>24,663</td>
<td>36,920</td>
</tr>
<tr>
<td>Crimes Against Society(^1)</td>
<td></td>
<td>634,151</td>
</tr>
<tr>
<td>Drug/Narcotic Offenses</td>
<td></td>
<td>550,656</td>
</tr>
<tr>
<td>Gambling Offenses</td>
<td></td>
<td>1,238</td>
</tr>
<tr>
<td>Pornography/Obscene Material</td>
<td></td>
<td>5,795</td>
</tr>
<tr>
<td>Prostitution Offenses</td>
<td></td>
<td>10,215</td>
</tr>
<tr>
<td>Weapon Law Violations</td>
<td></td>
<td>66,247</td>
</tr>
</tbody>
</table>

Table 6: Number of victims by type of offense in final NIBRS 2011 analysis file

\(^1\) Crimes against society by definition cannot have an individual as a victim.
Some analyses in the following chapters are restricted to counties in the Midwest only, defined by the Census Bureau as the states of Illinois, Indiana, Iowa, Kansas, Minnesota, North Dakota, South Dakota, Missouri, Nebraska, Ohio, Wisconsin, and Michigan. The restriction to the Midwest is an attempt to remove some of the bias due to agencies that do not report to NIBRS.

Ideally, we could assume that the missing agencies are missing completely at random (MCAR)—that is, the agencies that do participate in NIBRS are approximately a random sample of all agencies. We know that this is not the case, however. There is a recognized medium-agency bias, where medium-sized agencies are more likely to participate in NIBRS than large or small agencies. It is thought that large agencies have large enough budgets that the incentives to join NIBRS (enhanced reporting capabilities, assistance with setting up the new reporting system) do not offset the burden of changing their current systems, and small agencies do not report enough crimes to justify the administrative burden of switching to NIBRS. NIBRS status also depends heavily on the state's decision to adopt NIBRS, which is seen clearly in Figure 1.

When looking at the number of NIBRS and non-NIBRS agencies by MSA status, population grouping, and region as shown in Table 7, NIBRS coverage in the Midwest appears relatively high and relatively evenly spread among all groups. (Table 7 was created using the agency-level 2011 NIBRS data file.) The shaded rows indicate groupings where 50% of the agencies or more participate in NIBRS. Nearly all such rows
are in the Midwest. Each size by MSA category in the Midwest has somewhere between 50-60% NIBRs participation, with the exception of large non-MSA counties, of which there are none in the Midwest. NIBRS participation rates in other regions are generally much lower, and vary considerably. The Northeast has particularly low rates, especially for MSA and non-MSA counties; these rates are low enough that any analyses relying on NIBRS data from the Northeast should be interpreted with extreme caution.
<table>
<thead>
<tr>
<th>Population group</th>
<th>Cities</th>
<th>MSA Counties</th>
<th>Non-MSA Counties</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-NIBRS</td>
<td>NIBRS</td>
<td>% NIBRS</td>
</tr>
<tr>
<td><strong>North-east</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Under 25,000</td>
<td>2652</td>
<td>535</td>
<td>16.8%</td>
</tr>
<tr>
<td>25,000-99,999</td>
<td>164</td>
<td>86</td>
<td>34.4%</td>
</tr>
<tr>
<td>100,000 &amp; over</td>
<td>23</td>
<td>8</td>
<td>25.8%</td>
</tr>
<tr>
<td>Overall</td>
<td>2839</td>
<td>629</td>
<td>18.1%</td>
</tr>
<tr>
<td><strong>Midwest</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Under 25,000</td>
<td>1918</td>
<td>2198</td>
<td>53.4%</td>
</tr>
<tr>
<td>25,000-99,999</td>
<td>99</td>
<td>150</td>
<td>60.2%</td>
</tr>
<tr>
<td>100,000 &amp; over</td>
<td>20</td>
<td>28</td>
<td>58.3%</td>
</tr>
<tr>
<td>Overall</td>
<td>2037</td>
<td>2376</td>
<td>53.8%</td>
</tr>
<tr>
<td><strong>South</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Under 25,000</td>
<td>2947</td>
<td>1937</td>
<td>39.7%</td>
</tr>
<tr>
<td>25,000-99,999</td>
<td>126</td>
<td>94</td>
<td>42.7%</td>
</tr>
<tr>
<td>100,000 &amp; over</td>
<td>67</td>
<td>30</td>
<td>30.9%</td>
</tr>
<tr>
<td>Overall</td>
<td>3140</td>
<td>2061</td>
<td>39.6%</td>
</tr>
<tr>
<td><strong>West</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Under 25,000</td>
<td>976</td>
<td>503</td>
<td>34.0%</td>
</tr>
<tr>
<td>25,000-99,999</td>
<td>142</td>
<td>46</td>
<td>24.5%</td>
</tr>
<tr>
<td>100,000 &amp; over</td>
<td>93</td>
<td>20</td>
<td>17.7%</td>
</tr>
<tr>
<td>Overall</td>
<td>1211</td>
<td>569</td>
<td>32.0%</td>
</tr>
</tbody>
</table>

Table 7: 2011 agency-level NIBRS coverage, by region, MSA status, and population grouping
The American Community Survey (ACS) was designed to provide timely, detailed information about the U.S. population and selected social, economic, and housing trends, with an emphasis on providing estimates for small areas like counties, census tracts, or school districts. The ACS questionnaire asks for demographic information about the persons living in the housing unit such as name, age, sex, and race, as well as specific population data like citizenship, ancestry, health insurance status, and employment status. It also collects information about sampled housing units, such as the year built, number of units in structure, number of rooms, and access to services like plumbing or Internet access. It replaces the long-form U.S. Census questionnaire, which was sent to about one out of every six addresses in Census 2000 and prior censuses. The long-form data provided a snapshot of the U.S. on census day only once every ten years; updated ACS data are released every year, starting with 2005. To control the variability of published estimates, the Census Bureau releases estimates averaging five years of ACS data (“5-year estimates”). Places with more than 20,000 people have 3-year ACS estimates as well as the 5-year estimates, and single-year estimates are available for large places (populations of 65,000 or more).

We do not use the ACS data extensively, so we provide only a brief description of the datasets used. For more information on ACS methodology, see Torrieri et al. 2014, from which much of the information in this section is taken. The ACS is a housing unit (HU)
based survey that samples approximately three million residential addresses each year, with an additional sample of 20,000 group quarters in the U.S. and 36,000 residential addresses in Puerto Rico. A sample is taken from each U.S. county in every year.

Addresses are drawn from the Census Bureau’s Master Address File (MAF), which is constantly updated to be as complete as possible. Before the 2005 ACS sample selection, the MAF was divided into five sub-frames, each containing 20% of the addresses in each U.S. county. In the following years, new addresses are systematically allocated to each of the five sub-frames. Each sub-frame is assigned to a year, and the sub-frames are rotated annually to fulfill the requirement that no housing unit (HU) can be sampled more than once in a five-year period. For example, the sub-frame of addresses used this year (2015) will be used again in 2020 (updated as necessary). Addresses are sampled from the selected sub-frame by county according to sampling rates set for each of 16 strata. Thirteen sampling rates vary year-to-year; three are fixed. In 2013, the sampling rates ranged from 15% of addresses in the smallest areas to 0.5% of addresses in the largest areas. This yearly sample is then allocated into twelve months of data collection.

Sampled households with a mailable address are first sent a request with a link to complete the questionnaire online, followed by a paper survey if the household fails to respond within a month. Nonrespondents to the paper survey with a valid telephone number receive CATI follow-up. If the household fails to respond to the phone interview, fails to respond to the mail survey but does not have an associated telephone number, or does not have a mailable address, it may be selected for follow-up via personal interview.
CAPI sampling rates range from 33% in the largest areas to 100% in the smallest areas. The ACS is a mandatory survey, so response rates are typically extremely high—in 2012, 97% of sampled households completed an interview (Torrieri et al. 2014).

The ACS is the most reliable source for county-level population data, and ACS estimates are frequently used as fact in population research, ignoring any variance. ACS tables can be generated through the American FactFinder website at http://factfinder.census.gov/, and public-use microdata (PUMS) files are also available. In Chapter 4, we use ACS data to estimate the distribution of county populations in 2011 by age category, sex, and race. To do so, we used the American FactFinder website to find two county-level tables: table B01001, which tabulates the number of persons in each sex by age category, and table B01001A, which tabulates the number of white persons in each sex by age category. (A county-level table of age by sex by race was not available). We used the 5-year estimates, since some counties of interest have less than 20,000 persons and do not have 1-year or 3-year estimates.

For each county, we excluded persons less than 10 years old to match up with the NCVS data as closely as possible; the age categories available in the ACS data do not have a cut-point at exactly 12 years. Table B01001 provided the total number of persons ages 10 and older living in each county, which we used as the population size. We used table B01001A to calculate the percentages in each sex by age (10-29, 30-54, 55+) category for whites, then subtracted the values in table B01001A from the values in table B01001 to
calculate the percentages for nonwhites. For example, we could use table B01001 to find that there were 10,000 total persons 10 and older living in county A, and 2,000 of them were 10-29 year old males. Table B01001A could tell us that there were 1,500 white males 10-29 years old living in county A. Then we would calculate that 1,500/10,000= 15% of the 10 and older population of county A is white, male, and 10-29 years old, and (2,000-1,500)/10,000= 5% is nonwhite, male, and 10-29 years old.

Figure 2 shows the distribution of the ACS population percentages for each of the 1,452 counties that have usable NIBRS data. Notice the large number of counties with zero or near-zero percentages for the non-white categories.
Figure 2: Distribution of ACS population percentages for counties with NIBRS data (n=1452)
Addington and Lynch (2007) observe in the introduction to their thorough discussion of UCR and NCVS data that most researchers have chosen to use either UCR or NCVS data because it is simply too difficult to combine the two sources in a meaningful way. They cover different populations, define crimes differently, count crimes differently, and were designed for two very different purposes. NIBRS poses some of the same challenges as the UCR, but because it provides much more information on each incident it is possible to subset the data in a way that agrees much more closely with the NCVS data.

The NCVS is designed to estimate the total number and types of victimizations perpetrated against non-institutionalized persons ages 12 and older in the United States. It is specifically designed to capture crimes not reported to the police, and excludes crimes committed against non-individuals such as bank robberies. The UCR summary counts only cover crimes reported to the police from agencies that choose to report to the FBI, but there is no way to tell the ages of victims, or even if the victim is a business or an individual. NIBRS does include the age of the victim and also an indicator for whether the victim was an individual or a commercial establishment, so it is possible to subset the NIBRS file to include only individual victims 12 and older to match the population of the NCVS.
NIBRS crime rates are given in terms of incidents per person. Since the NCVS explicitly intends to estimate victimization rate, NCVS crime rates are reported in terms of victimizations per person. Using the hotel example from section (b), if five people have their hotel rooms burglarized, NIBRS/UCR reports one incident, while the NCVS reports five victimizations. The NCVS does provide incident-level weights, so it is possible one could reduce the NCVS data to the incident level. However, crime rates are arguably more intuitive when they reflect number of victimizations rather than incidents—two people being held up at gunpoint in one incident should affect the crime rate more than if only one person is robbed. Therefore, we used the victim-level information in NIBRS to count how many people were victimized in each incident of interest.

A serious concern when combining NIBRS data and NCVS data is that the two sources often use very different definitions of crime. NIBRS relies on the reporting officer to determine what type of crime occurred and code it properly. Definitions of crime can vary across jurisdictions and years, although to report through NIBRS agencies must use definitions that are reasonably consistent with the FBI definitions. The NCVS, on the other hand, takes a description of each victimization from the respondent and BJS staff later assigns an incident type based on a list of standardized definitions. Rape, for example, is historically one of the most difficult crimes to compare between NIBRS and NCVS. NCVS rape rates include attempted rape of any kind, any unwanted sexual contact that may or may not involve force, and statutory rape, perpetrated against either
male or female victims. NIBRS rape includes only forcible rape, although the FBI definition has recently been expanded to cover both male and female victims.

We also want to consider how frequently a crime is reported to the police. Information on police-reported crime is available from the NCVS and from NIBRS; information on crime not reported to the police can only be obtained from the NCVS. In fact, since NIBRS is designed as a census of police-reported crime, we only have to estimate the rate of crimes not reported to the police. If the percentage of crimes not reported to the police is relatively small, then our estimates of overall crime rates will be more accurate simply because we are using more known information. We would prefer to use a crime type with a reasonably high reporting rate as we develop our methodology. The model could then be extended to crimes with lower reporting rates.

The property crimes (burglary, theft, motor vehicle theft) tend to have the most similar definitions between NIBRS and the NCVS and have historically high reporting rates, but the NCVS considers those household-level crimes. This means that we would not be able to assign victim-level characteristics to those crimes; we could use the demographic characteristics and personal weight of the reference person in the household as a proxy, but this would add another level of approximation. Simple assault is the personal crime type with the largest number of reports; however, simple assault is frequently not reported to the police. In 2011, BJS estimated that only about 45% of simple assaults were reported to the police (Planty & Truman 2011), likely because people often don't
think of simple assaults as a crime, like a shoving match between friends or a punch thrown in a bar at the end of the night. Notice that in Table 3, even “verbal threat of assault” is also included under simple assault in the NCVS. Simple assault is very difficult to measure accurately, and is likely underreported even in the NCVS; victims may forget a fight with a friend or a verbal threat that happened four or five months before the interview, or they may not think it worth reporting.

Addington (2007) suggests that of the personal crimes, the definitions of aggravated assault correspond the most closely across the two sources. NIBRS uses the FBI definition of aggravated assault as “…an unlawful attack by one person upon another for the purpose of inflicting severe or aggravated bodily injury… usually accompanied by the use of a weapon or by other means likely to produce death or great bodily harm” (NIBRS 2011). The NCVS definition is “… an attack or attempted attack with a weapon, regardless of whether or not an injury occurred, and attack without a weapon when serious injury results. Serious injury includes broken bones, loss of teeth, internal injuries, loss of consciousness, and any injury requiring two or more days of hospitalization” (NCVS 2011). The police do have some latitude in NIBRS to decide whether to classify an assault as simple or aggravated when a weapon is not used, but the NCVS definition describes such serious injuries that an NCVS aggravated assault victimization would likely be classified as an aggravated assault by the police as well, even if a weapon is not used. Although the NCVS definition includes attempted assaults with a weapon while the NIBRS definition does not seem to, NIBRS contains one
variable for offense type and a second variable for whether the offense was attempted or completed. We include both attempted and completed assaults from both data sources for consistency. From this point on, the term “aggravated assault” will refer to both attempted and completed aggravated assaults.

Aggravated assault is also an offense that is violent enough that it is widely reported; Criminal Victimization 2011 estimates that about 70% of aggravated assaults were reported to the police in collection year 2011, and Table 3 in Section 2.1 shows a 66% reporting rate for aggravated assault in data year 2011. It also occurs frequently enough that we can analyze the data with some level of detail. There are 246 reported aggravated assault victimizations in the 2011 NCVS data year file, which translates to a weighted rate of about 4.0 aggravated assaults per 1000 persons. While not a huge number, it is possible to create an age by sex by race table of aggravated assault cases with no empty cells, as in Table 8. In contrast, in 2011 the estimated rate of robbery was 1.0 per 1,000 persons, with only 30% of robberies reported to the police—this corresponds to only 118 reported robbery victimizations in the entire 2011 NCVS data year file.

<table>
<thead>
<tr>
<th>Age</th>
<th>12-29</th>
<th>30-55</th>
<th>55+</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Non-white</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>14</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>Female</td>
<td>17</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td><strong>White</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>54</td>
<td>45</td>
<td>11</td>
</tr>
<tr>
<td>Female</td>
<td>29</td>
<td>32</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 8: Count of aggravated assaults in the 2011 NCVS by age, sex, and race of victim
Limiting our analysis to the crime of aggravated assault should simplify the model-building process, but estimating county-level crime rates for aggravated assault is still a difficult and complex problem. Chapter 3 summarizes several recent papers attempting to estimate small-area crime rates, all of which are either largely unsuccessful or focus only on the police-reported crime rate; we also briefly review some of our own exploratory analyses. Chapters 4 and 5 present two strategies for estimating county-level crime rates, one simulation-based and one model-based, applied to the crime of aggravated assault.
Chapter 3: Previous and related work

Section 3.1: Review of relevant literature

Most of the published work on combining police crime data with the NCVS attempts to use UCR data as auxiliary information in a regression model, and typically finds that UCR crime rates are not good predictors of the corresponding NCVS crime rates (Fay & Li 2011; Fay & Diallo 2012; Rosenfeld 2007). We suspect this is largely because the UCR summary counts are such a blunt metric; they can be broken down geographically, but not by demographic characteristics. The tactic that is often taken is to build a regression model using UCR crime rate to predict NCVS crime rate, and including other predictors like poverty rate or proportion renter as a proxy for rate of unreported crime. UCR personal crime rates are generally not good predictors of the corresponding NCVS rates because of differences between the definitions of each crime type and the populations covered, as described in greater detail in Chapter 2. It is also possible to access BJS's proprietary data with county identifiers, but Fay and Li (2011) found that they were not able to make good predictions for county-level crime rates when they tried to predict NCVS violent crime rates using UCR violent crime rates via linear regression. Using NIBRS' greater level of detail should help resolve some of these issues.
The problem of estimating county-level crime rates falls under the broad category of small-area estimation. The classic small-area estimation problem occurs when we are interested in a quantity at a fine level of detail (perhaps insurance rates for people by age, sex, race, and income, or county-level poverty rates, for example), but there is not enough information available at that level to produce useful estimates. In many applications, there are some small domains of interest in which no information is available—for example, a county in which no one was sampled. However, stable estimates are usually available at some higher level, such as at the state or region. Small-area estimation techniques are generally based on the idea of “borrowing strength” across small domains or from some aggregation to a higher level to improve estimation by using some type of indirect estimator. For an in-depth introduction to small area estimation, see J.N.K. Rao's excellent text *Small Area Estimation* (2003). Rao provides a comprehensive overview of small area estimation theory and explains a wide range of possible models in detail, using examples from published papers.

Fay and Herriot's 1979 paper is widely considered the foundation of small-area estimation theory. Fay and Herriot were trying to create estimates of per capita income (PCI) for approximately 39,000 places in the United States based on data from the 1970 Census of Population and Housing as part of their work at the U.S. Census Bureau. A place is defined as some unit of local government; places range in size from very large cities all the way down to townships and unincorporated places. The sampling error was unacceptably large for small places, so the Census Bureau somewhat arbitrarily decided
that for places of under 500 persons the county-level estimate of PCI should be used rather than the place-level estimate. This reduced sampling error, but discarded useful information about mean PCI of small places. As a result, the potential bias could be severe if the small place was very different from the rest of the county.

Fay and Herriot proposed a method based on the James-Stein estimator, a shrinkage estimator for multivariate problems which uniformly improves estimation under squared error loss by combining information across (even unrelated) problems. To estimate PCI for county i, $Y_i$, assume:

$$Y_i|\theta_i \sim N(\theta_i, \psi_i), \quad \theta_i|\beta, \sigma^2 \sim N(X_i\beta, \sigma^2) \quad (3.1)$$

for $X_i$ a p-dimensional vector of covariates for small area i and regression coefficients $\beta$ with an improper uniform prior. The $Y_i$s are assumed conditionally independent given $\theta_i$, and similarly assume that the $\theta_i$s are conditionally independent given $\beta$ and $\sigma^2$. This model assumes that the covariates $X_i$ and the sampling variances $\psi_i$ are known, but that $\beta$ and $\sigma^2$ would need to be estimated from the data. $\theta_i$, the true county-level mean, is the quantity of interest but is not observed directly—only $Y_i$ is observed. The first equation is frequently referred to as the *sampling model*, because it describes how observations are sampled from the small area population; the second equation is the *linking model*, because it describes the link between the small area parameter $\theta_i$ and the selected covariates.

Fay and Herriot show that the James-Stein estimator for $\theta_i$, $\tilde{Y}_{iFH}$, in this case is given by
\[ \bar{Y}_{iFH} = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \psi_i} y_i + \frac{\psi_i}{\hat{\sigma}^2 + \psi_i} \hat{Y}_i \]  

(3.2)

where \( \hat{\sigma}^2 \) is the sample estimate of \( \sigma^2 \), \( y_i \) is the direct estimate, and \( \hat{Y}_i \) is the regression estimate of \( Y_i \). Thus, \( \bar{Y}_{iFH} \) is essentially a weighted average of the sample and regression estimates. In practice, the sampling variances \( \psi_i \) are typically unknown and must be estimated as well.

Fay and Herriot fit a regression model to predict log PCI for each place using log county PCI, log average adjusted gross income per exemption for both place and county from 1969 IRS data, and log total value of owner-occupied nonfarm housing from the 1970 Census for both place and county. Their analysis was restricted to places of under 1,000 people, because the relationship between the covariates and log PCI may be different for larger places. Separate models were fit for places with fewer than 500 people and places with 500-999 people by state for a total of 100 regressions, 50 states by two size categories. Large values of \( \hat{\sigma}^2 \), the estimated variance of \( \theta_i \), indicated a poor model fit and would increase the weight placed on the direct sample estimate. Finally, Fay and Herriot constrained all estimates to be within one standard error of the sample estimate, so that a regression estimate from a poorly fitting model could not influence the final estimate too heavily. They found that in nearly all cases, their estimator outperformed using sample values alone or using the current Census Bureau procedure.
Although some county-level information is used, Fay and Herriot's method largely assumes complete data at the small-area level, which is not always the case—for the NCVS, data is not publicly available at the county level. Even with access to NCVS county-level identifiers, many counties would have no NCVS sample households, so most Fay-Herriot estimates would be solely based on the regression estimates.

Elliott and Davis (2005) provide a solution for the case where some auxiliary information is available only at a higher level. They want to estimate cancer risk factors at the county level. Researchers typically use either the National Health Interview Survey (NHIS) or the Behavioral Risk Factors Surveillance System (BRFSS) to estimate risk factor prevalence at the national or state level, but neither survey is designed to produce reliable small area estimates. The NHIS is conducted directly by the National Center for Health Statistics in about 40,000 households each year and is generally considered to provide high-quality data: it uses a face-to-face interview with a high response rate and well-trained interviewers. The BRFSS is larger (about 160,000 persons each year) and covers far more counties. However, the BRFSS is administered by individual states, so there is variation in the quality of survey administration. It is also a land-line only telephone survey, with a much lower response rate than the NHIS (47.8% in 2000, as compared to 88.9% for the NHIS)—both of which can lead to a considerable middle-class bias, undercounting both low- and high-income groups.
This is very similar to the relationship between the NCVS and NIBRS data—one data source (NIBRS/ BRFSS) is more widely available at the county level, but the information is likely biased and varies in quality depending on the locality or state, while the other data source (NCVS/NHIS) is a nationally administered survey that produces high-quality data, but that data is not publicly available at the county level and is too sparse to be useful even with county identifiers. Elliott and Davis present their method as one that “combine[s] data from two or more population-based surveys where one survey…is assumed to be 'least biased' and one or more surveys in overlapping areas…to have larger sample sizes but may suffer from a larger degree of frame or inclusion bias… [with] the additional complication that small-area identifiers are not generally available to the public” (p. 597).

Elliott and Davis use propensity scores to adjust the BRFSS weights to match the distribution of covariates in the NHIS, selecting all common covariates between the two surveys to correct for as much bias as possible\(^5\). A propensity score is calculated for each individual in each survey as the odds of a subject in a given large area being in the NHIS (rather than the BRFSS), conditional on the outcome of interest and the covariates relative to the odds of a subject being in the NHIS given the large area only:

\[
s_i = \frac{P(\text{in NHIS} | Y_i, X_i, A_i = r) / P(\text{in NHIS} | G_i = g)}{P(\text{in BRFSS} | Y_i, X_i, A = r) / P(\text{in BRFSS} | G_i = g)}
\]

\(^5\) These are: age, general health status (five points, excellent to poor), education, income, race/ethnicity, marital status, and health insurance status.
where $Y_i$ is the outcome of interest, $X_i$ is the vector of covariates, and the $G_i$ are all counties within large area $A_i$ for individual $i$. The probabilities of being in the NHIS or in the BRFSS are calculated based on the total number of respondents to the NHIS and the BRFSS, not from the entire population of the United States. The authors suggest using as covariates all common measures that may be associated with the outcome of interest, or for which estimates differ widely between the two surveys. Large areas $A_i$ are defined by region and MSA population categories, for a total of 27 nonempty cells. The small areas of interest are counties $G_j$ in which there are at least 50 BRFSS respondents; each county will be nested in some large area.

Elliott and Davis find that the propensity score adjustment tends to increase variance; in some cases, the reduced bias from incorporating the NHIS data is not enough to offset the increased variance, leading to larger mean squared error (MSE). The final proposed estimator is a hybrid estimator, in which the adjusted estimator is chosen only when its MSE is smaller than that of the unadjusted estimator. The authors look at the outcomes male smoking prevalence and female mammogram usage, using 1999-2000 BRFSS and NHIS data. The hybrid estimator estimated male smoking prevalence to be 1.1 percentage points higher and female mammogram usage to be 4.9 percentage points lower than the unadjusted BRFSS data—both directions which would correct for the bias the researchers expected to find. However, the median county-level standard errors also increased by 25-30% for each outcome.
Elliott and Davis’ propensity score method assumes that both sources of data used have survey weights that can be adjusted. NIBRS is not a survey, and therefore there are no NIBRS survey weights to adjust. There is no probability that a person will be sampled in NIBRS—an individual's probability of being in NIBRS depends on the probability of victimization and the probability that the individual will file a police report, as well as on whether or not the covering agency reports to NIBRS. While we know what agencies report to NIBRS, we do not know the other two probabilities. Elliott and Davis' method also relies on individual-level data from respondents both with and without the outcome characteristic of interest; for example, both surveys contain responses from both smokers and non-smokers. NIBRS only collects individual-level information on police-reported victimizations, and contains no individual-level information on persons who were not victimized. Propensity score adjustments further require that all elements in the population have a probability of selection that is not exactly zero or one—this is not the case in the NIBRS file, since people who reported at least one crime committed against them in 2011 are selected with probability one, and all others are selected with probability zero.

Lohr and Brick (2011) also address the issue of combining data from two surveys for small-area estimation, with one of the surveys being the NCVS. In summer 2012, BJS pilot-tested a Companion Survey (CS) to the NCVS. The CS was designed to field additional sample in selected small domains to improve accuracy for small domain
estimates without increasing the sample size of the full NCVS. BJS designed the CS to be cheaper to administer than the full NCVS. It uses address-based sampling rather than the multi-frame sample of the full NCVS, asks adults about victimizations over the past 12 months (rather than 6), and consists of only one telephone or mail interview per household. Because of these differences, it is expected that the CS crime rates will be biased relative to the NCVS rates, but the direction and mechanism of the bias is unclear. The CS will likely have a much lower response rate, and the NCVS has demonstrated that nonrespondents tend to be from groups that report more victimizations—young, unmarried, male, and nonwhite. However, the 12 month recall period and the single survey may lead to more reported victimizations. In the NCVS, respondents often report fewer victimizations in later waves, because they have learned that reporting a crime leads to a long series of questions; other household members might also be aware that a long survey means that someone is reporting a crime, discouraging respondents from reporting domestic victimizations or anything they have not shared with their family.

As of the writing of their paper Lohr and Brick did not have the CS data available, so they set up a series of simulations showing how several different methods of combining estimates would perform under different scenarios. The scenarios they considered were no relative bias between the surveys, constant additive bias, constant multiplicative bias, or differential additive or multiplicative bias across the small domains of interest. They then considered four general models:
1) Weighted average, same weight across all domains: \( \lambda \hat{Y}_d + (1 - \lambda) \hat{X}_d \)

2) Weighted average, weight may vary across domains: \( \lambda_d \hat{Y}_d + (1 - \lambda_d) \hat{X}_d \)

3) Additive bias model: \( \lambda_d \hat{Y}_d + (1 - \lambda_d)(\hat{X}_d - \text{est } bias_d) \)

4) Multiplicative bias model: \( \lambda_d \hat{Y}_d + (1 - \lambda_d) b \hat{X}_d \)

In each case, \( \hat{Y}_d \) is the estimate of victimization rate in domain d based on NCVS data only, and \( \hat{X}_d \) is the estimate based on CS data only. In the additive bias model, \( \text{est } bias_d \) is an estimate of the bias in domain d. For their simulation, the authors did not have access to the county-level NCVS data so they considered demographic subgroups as domains: race, age group, rent/own, and number of times household has moved. The outcomes of interest were violent crime rate and property crime rate in each domain. In each case, Lohr and Brick assumed that the national CS estimates were calibrated to the national NCVS estimates, suggesting that it would be best to calibrate the household and person weights by crime type so that the associated totals from the CS summed to the national NCVS totals. The calibrated CS could not be used to improve national NCVS estimates, but would likely introduce less bias at the small area level.

Lohr and Brick found that under the case with no simulated bias, incorporating the CS information uniformly improved MSE over using the NCVS data only, with the models that assume no bias performing best. When bias was introduced, not surprisingly the models assuming multiplicative bias performed best under multiplicative bias, and the models assuming additive bias performed best under the additive bias condition. Almost all models assuming bias outperformed using the NCVS information only. However,
under differential bias, only one of the additive bias models and the multiplicative bias model reduced MSE—the authors conclude that the multiplicative bias model “appears to reduce MSE at best and do no harm at worst.”

This paper, like Elliott and Davis', also demonstrates that model-based estimators can often increase MSE; Lohr and Brick did not use a weighted average of the estimate and the survey value in the final step, but it seems possible that could have reduced some of the problems they had under the scenarios where bias differed across domains. However, much of their work again depends on both data sources containing information on the same outcome variables at the individual level. Even the idea of calibrating the national estimates between the two surveys, which according to Lohr and Brick nearly uniformly reduces bias, cannot be used for our research because NIBRS does not have complete national coverage and it is impractical to restrict the NCVS data file to only sample cases in NIBRS counties.

Sharon Lohr also published a related 2008 paper with one of her students, Lynn Ybarra, which directly addresses the weighted average step of the estimator (Ybarra and Lohr 2008). Ybarra and Lohr note that the Fay-Herriot estimator has been shown to reduce MSE in most cases, but only under the assumption that the additional information used is unbiased or that it is possible to adjust for any bias—the assumption is that incorporating the auxiliary information will reduce bias enough to make up for the increase in variance. If the auxiliary information has substantial measurement error, we may not be able to
correct for any bias and the Fay-Herriot estimator may have increased bias and variance over the direct estimator. Ybarra and Lohr refer to Elliott and Davis' 2005 paper, saying that the estimators presented in that paper are conditionally biased given the true auxiliary variables because the NHIS estimates have nonsampling error that is not accounted for.

The Fay-Herriot model can be written as

\[ y_i = X'_i \beta + v_i + e_i, \quad i = 1, 2, ..., n \]  

where \( v_i \) is the model error and \( e_i \) is the design error, both independent mean 0 random variables. Under the typical assumption that \( v_i \sim N(0, \sigma_v^2) \) and \( e_i \sim N(0, \psi_i) \), with \( v_i \) and \( e_i \) mutually independent, the best unbiased linear predictor (BLUP) of \( Y_i \) is:

\[ \hat{Y}_{iFH} = \frac{\sigma_v^2}{\sigma_v^2 + \psi_i} y_i + \frac{\psi_i}{\sigma_v^2 + \psi_i} X'_i \beta \]  

(3.5)

with MSE given by \( \frac{\sigma_v^2 \psi_i}{\sigma_v^2 + \psi_i} \). However, typically \( \beta \) and \( \sigma_v^2 \) are unknown and must also be estimated; plugging in their estimates will give the empirical BLUP. (Again, the \( \psi_i \) are usually considered known, even though in practice they must be estimated.)

If the covariates \( X \) are not known, then we need to use the estimator \( \hat{X}_i \) which has an associated MSE—let \( \text{MSE}(\hat{X}_i) = C_i \). (In practice, \( \text{MSE}(\hat{X}_i) \) is also not known and must
This MSE must be incorporated in the MSE of the Fay-Herriot estimator, giving

\[
MSE(\bar{Y}_{iFH}) = \frac{\sigma_v^2 \psi_i}{\sigma_v^2 + \psi_i} + \left(\frac{\psi_i}{\sigma_v^2 + \psi_i}\right)^2 \beta' C_i \beta. \tag{3.6}
\]

Ybarra and Lohr point out that if \(\beta' C_i \beta > \sigma_v^2 + \psi_i\), then the MSE of this estimator is greater than the MSE of the direct estimator using the \(y\) information only. They propose the alternative estimator \(\bar{Y}_{IME}\), where ME stands for “minimum error”:

\[
\bar{Y}_{IME} = \frac{\sigma_v^2 + \beta' C_i \beta}{\sigma_v^2 + \beta' C_i \beta + \psi_i} y_i + \frac{\psi_i}{\sigma_v^2 + \beta' C_i \beta + \psi_i} X_i' \beta \tag{3.7}
\]

and show that this estimator has minimum MSE. They also suggest consistent estimators of \(\sigma_v^2\) and \(\beta\) that minimize the MSE of the empirical estimator. A simulation study comparing this estimator to the standard Fay-Herriot estimator showed that the Fay-Herriot estimator consistently underestimated the empirical MSE when the auxiliary information \(X_i'\) contained random error. The authors also tested their estimator by looking at BMI in 50 demographic subgroups as reported from the National Health and Nutrition Examination Survey (NHANES) and the BRFSS in 2004. The BMI measurement from NHANES is measured by a medical professional and can be considered extremely accurate; the BRFSS simply asks respondents to self-report BMI. Here, NHANES is the high-quality survey with less data (4424 valid responses), and the BRFSS data is assumed
to be measured with error, but there is much more of it (29,652 responses). Ybarra and Lohr found that incorporating the auxiliary information from the BRFSS improves MSE in nearly all subgroups.

All of the papers mentioned above focus on applications where all outcome data were available at the individual level, but the NIBRS data only allows for a county-level crime rate. A 2008 paper by Calder et al., “Relating Ambient Particulate Matter Concentration Levels to Mortality Using an Exposure Simulator,” provides a possible strategy for linking the personal-level NCVS data to the county-level NIBRS data. High concentrations of small airborne particles in the atmosphere, called particulate matter (PM), have been correlated with an increase in health problems, such as asthma or cardiovascular issues—smog is a striking example of high PM levels. Most research on the effect of PM has been done by taking a PM reading at an air quality station and using a statistical model to relate the effect of that concentration to some geographic area around that station. However, these types of models do not accurately reflect a person's daily PM exposure. Even if the ambient PM level is very high, a person who stays indoors all day will not be exposed to that level of PM; their exposure will depend much more heavily on the ambient PM concentration inside the home.

Calder and her co-authors focused on the effect of particles less than 2.5 nanometers in diameter (PM$_{2.5}$) on deaths due to cardiovascular causes in eight counties in central North Carolina. We will not explain the entire model in detail, only the portions of interest to
our research—the county-level portion of the model involves building a latent spatial field of ambient PM$_{2.5}$ levels, and the final step of the model links average exposure to number of cardiovascular deaths via a Poisson regression model. To estimate daily PM exposure of persons in the counties of interest, Calder et al. used data from the National Human Activity Pattern Survey (NHAPS); NHAPS is a national survey which asks participants to keep a diary of their activities throughout the day. First, they divided the population of each county into a table based on sex, age, and employment status from the summary files of the 2000 U.S. Census. They then sampled 100 individuals from the NHAPS according to the frequency distribution of this table. Although the selected persons could have come from anywhere in the country, by matching on sex, age, and employment status the authors argue that the selected activity patterns will provide a good representation for the true activity patterns in the study region. This approach is applied to the NCVS data in Chapter 4 to simulate the percentage of crimes reported to the police in each county.

Once Calder et al. select 100 individuals, the authors use a formula to calculate each individual's PM$_{2.5}$ exposure on a given day, based on the ambient outdoor concentration and the assumed ambient indoor concentration, adjusted by the amount of time spent cooking or smoking (both activities which release additional PM). The individuals' levels are then averaged together to generate one average county-level PM$_{2.5}$ exposure for county $c$ on given day $t$, $\tilde{\xi}_{ct}$. Finally, the authors assume that the true average exposure for a county is normally distributed, with mean $\tilde{\xi}_{ct}$ and unknown standard deviation. This
paper was able to link county-level data (daily number of cardiovascular deaths and PM$_{2.5}$ concentration) with individual-level data from a national survey of people outside the counties of interest by using a weighted average based on Census Bureau demographic information—a link that is missing in most other small area research. We use a similar strategy in Chapter 4 to link individual-level data from the NCVS to county-level data from NIBRS.

Williamson, Birkin and Rees (1998) address the problem of combining data at two different levels as well, but their focus is on estimating tables rather than parameters. The United Kingdom Census releases selected cross-tabulations down to the enumeration district (ED) level through the Small Area Statistics (SAS) program. However, not all possible combinations of Census variables are available through SAS. Microdata samples of Census records can be used to generate custom tables, but geographic information is available only at the regional level for these 1% Samples of Anonymised Records (SARs). The authors use what they term a combinatorial optimization approach to estimate custom tables at the ED level. The available SAS tables for a given ED are treated as constraints, and the goal is to find the sample of households from the 215,789 household records in the Household SAR file that best fit these constraints. For example, the authors illustrate their methods with data from ED DAFJ01, which contains 199 households and 532 persons according to the published SAS tables. The goal is to draw a sample of 199 households from the Household SAR file that approximates eight selected SAS tables for this ED as closely as possible. “Closeness” is defined as minimizing total
absolute error (TAE), which is simply the sum of the absolute differences between cell
counts in the observed (sample-generated) and expected (published) tables.

Williamson, Birkin and Rees used three general optimization algorithms. The first and
simplest is the hill climbing (HC) algorithm, in which an initial sample of 199 households
is randomly selected. At each step a new household is randomly selected with
replacement from all available households, and one household is randomly selected from
the households in the sample. If TAE is reduced by replacing the household in the sample
with the selected household, the household is replaced; otherwise, no change is made.
Because such an algorithm can get stuck in local minima, a variant is also used: instead
of randomly selecting a household to be replaced, select the household with the greatest
contribution to overall TAE for replacement. Each algorithm was tested with five runs of
500,000 steps each, with each run using a different initial sample of households. The
TAE of each run and the five-run average was reported. The second algorithm, simulated
annealing (SA), allows increases in TAE with a certain probability and is based on an
equation used to describe the cooling process in thermodynamics. Again, two variants are
tested, each with five runs of 500,000 steps. Finally, five variants of a genetic testing
algorithm are proposed, based on evolution toward the “optimal” chromosome via cross-
breeding and mutation. Five runs of 14,000 steps were carried out; the lower number of
steps is because these algorithms replace many households per step, while the HC and SA
algorithms replace only one household per step.
The simulated annealing algorithms tend to perform best, but not as well as the authors hoped. The authors also find that model fit depends heavily on how similar the enumeration district population is to the national population. If the table proportions are similar to national proportions, then fit tends to be good regardless of algorithm; if an ED is very different from the national populations, then the same uncommon households in the population tend to be added to the subsample multiple times and a good fit is difficult to achieve. The authors also suggest that more powerful computing could improve the results. This paper was first submitted in 1996, so better computing resources are certainly available. However, these methods are difficult to apply directly to NCVS/NIBRS data: the NCVS records are analogous to the microdata samples, but NIBRS tables would not be appropriate constraints because NIBRS only covers crime victims who report crimes to the police.

Other researchers have taken a hierarchical Bayesian approach to similar problems. Hierarchical Bayes models are attractive for small area estimation because they provide a convenient framework for incorporating information at different levels, and in a broad sense “weight” the information depending on its reliability. Ghosh and Rao (1994) note that empirical Bayes and hierarchical Bayes approaches generally seem to outperform other small area approaches such as synthetic estimators, composite estimators like the Fay-Herriot estimator, and non-Bayesian small area models. Farrell (2000) and Salvati et al. (2010) propose small area unit-level estimators based on hierarchical and empirical
Bayesian models; Yu, Stasny, and Li (2008) and Stasny (1991) build hierarchical Bayesian models for NCVS data, but do not use small area techniques.

Liu, Lahiri, and Kalton (2007) take a hierarchical Bayesian approach as well. They compare the performance of four hierarchical models for estimating small area proportions $P_i$. The first proposed model is the standard Fay-Herriot model, assuming a normal sampling model and a normal linking model:

$$Y_i|\theta_i \sim N(\theta_i, \psi_i), \quad \theta_i|\beta, \sigma^2 \sim N(X_i \beta, \sigma^2)$$ (3.8)

for $Y_i$ the survey weighted direct estimate of $\theta_i$, $X_i$ a p-dimensional vector of covariates and regression coefficients $\beta$ with an improper uniform prior. As before, assume the covariates $X_i$ and the sampling variances $\psi_i$ are known, and the $Y_i$ and $\theta_i$ are conditionally independent. The second model is the normal-logistic model, an extension of the Fay-Herriot model with a logit link function used in the linking model. The authors point out that both models assume the sampling variances $\psi_i$ are known, when in practice they are almost always estimated. In addition, the assumption of a normal sampling model may not be appropriate when the sample size in an area is small and the true mean small area proportion is close to 0 or 1, which is often true in small area applications. They therefore propose two additional models.
Model three is very similar to model two, but the sampling variances $\psi_i$ are considered unknown parameters approximated by the function

$$\psi_i = \frac{P_i \times (1 - P_i)}{n_i} \times def_iw$$  \hspace{1cm} (3.9)

where $def_iw$ is an approximation of the weighted design effect for small area $i$ and $n_i$ is the sample size in small area $i$. The fourth model modifies model three by using a Beta($a_i$, $b_i$) distribution for the sampling model, with the parameters given by

$$a_i = P_i \left( \frac{n_i}{def_iw} - 1 \right) \quad b_i = (1 - P_i) \left( \frac{n_i}{def_iw} - 1 \right).$$  \hspace{1cm} (3.10)

A simulation study was conducted using the 2002 Natality public-use data file, which contains data from all birth certificates filed in U.S. states and the District of Columbia in 2002. Liu, Lahiri, and Kalton used the 4,024,378 records of live births with birth weight data available to estimate $P_i$, the statewide proportion of births with low birth weight (under 2500 grams) for $i = 1, 2, \ldots, 51$. $P_i$ ranged from 5% to 11%. A stratified random sample of birth records was then drawn from each state using mother’s race (black, white, or other) as the stratification variable. The total national sample drawn was fixed at $n=4,526$ with state level sample sizes ranging from 7 (Vermont) to 690 (California). The sample was drawn a total of 1,000 times and $P_i$ was estimated for each sample using models 1-4. Any state-level direct estimates equal to 0 were replaced by a very small but
positive proportion for models 2-4 because the WinBUGS software used for model fitting cannot handle direct estimates of zero.

Estimates from model 1 tended to have very wide 95% credible intervals, averaging a width of 9.0%, but such wide intervals failed to cover the true parameters only 0.4% of the time on average. Model 2 had the narrowest credible intervals (5.5% average width), but its noncoverage rate tended to be high—on average, 8.2% of the true state-level parameters fell outside of the 95% credible intervals. Model 3 performed somewhat better, but its average noncoverage rate (6.5%) was also above the nominal noncoverage rate. Model 4 had a noncoverage rate closest to the nominal rate (4.4%), but at the cost of wider credible intervals (8.5% for model 4 vs. 6.2% for model 3). However, the authors conclude that model 4 appears the most promising—the wide intervals are probably due to extra variability stemming from issues with MCMC convergence when the survey-weighted proportions are close to zero. They also used no auxiliary variables, and note that including auxiliary variables would likely improve model fit for all four models.

You and Rao (2002) use a similar approach to perform sensitivity testing for different priors placed on models similar to model 1 and model 2. Esteban et al. (2011) also compare variations of hierarchical Bayes models to estimate the proportion of persons in poverty in Spanish provinces by sex. These two papers and the Liu, Lahiri, and Kalton paper, however, only consider data from one source and require that some data is available in each small area of interest, which is not true for the NCVS.
Wieczorek and Hawala, from the Census Bureau’s Small Area Income and Poverty Estimates program, also applied a hierarchical Bayesian model using a beta distribution in a 2011 paper extending the standard log rate area-level method using beta regression. In the log rate model, counties with ACS direct estimates of no children in poverty are currently dropped from the analysis, since log(0) is undefined. (Estimates are generated for these counties later by using their predicted values in the fitted model, but they are not used in the model fitting.) The authors also note that Census Bureau staff have concerns that the variance estimates may be biased under the log count model—one suggestion is to model rates rather than counts, as this seems to improve variance estimation while providing similar estimates for the county means. They suggest a hierarchical zero-one inflated beta regression model to model poverty rates; the beta allows them to model poverty rates directly, since its support is on the interval (0, 1), and the zero-one inflated component allows modeling of direct rates which are 0 or 1.

The model they propose is:

\[
\begin{align*}
    y_i | \mu_i, \gamma_i, x_i, \beta_\mu & \sim \text{Beta}(\mu_i, \gamma_i) \\
    \text{logit}(\mu_i) &= x_i' \beta_\mu \\
    p(\beta_\mu) & \propto 1
\end{align*}
\]  

(3.11)

\footnote{A beta distribution is traditionally parameterized as Beta(a, b), where the mean is equal to a/(a+b) and the variance is ab/(a + b)(a+b+1). However, it is often easier to think of the parameters under the reparameterization of the mean, \( \mu = a/(a+b) \), and a parameter related to the variance, \( \gamma = a + b \), with the variance then equal to \( \mu(1-\mu)/(\gamma+1) \). We use this parameterization here.}
where $y_i$ is the ACS estimate of county poverty rate, $\mu_i$ is the mean of the beta distribution for that county and $\gamma_i$ is a parameter related to the variance, $x_i$ is a vector of covariates, and $\beta_\mu$ is a vector of regression coefficients with an improper flat prior placed on it. To add the zero-one inflated component, they let $y_i$ be the true county poverty rate, and $Y_i$ be the observed ACS estimate of poverty rate in county $i$. Based on the true county poverty rate, there is some probability of observing a zero (call this $p_i^{(0)}$) and some probability of observing a one (call this $p_i^{(1)}$). Then, one can think of a multinomial trial to determine which outcome we actually observe:

$$Y_i = \begin{cases} 
0 \text{ w.p. } p_i^{(0)} \\
1 \text{ w.p. } p_i^{(1)} \\
\text{Beta}(\mu_i, \gamma_i) \text{ w.p. } 1 - p_i^{(0)} - p_i^{(1)}
\end{cases} \quad (3.12)$$

Wieczorek and Hawala use tax poverty rate, tax non-filing rate, food stamp participation rate, and natural log of the number of persons in poverty sampled by the ACS as covariates, the same covariates as in the standard SAIPE model. They used the 2009 SAIPE data to find the posterior estimates of the regression coefficients using MCMC, then used these coefficients to define the “true” values of all other parameters in the model and simulate 50 new datasets for testing. They find that their model does well at predicting zero rates for counties with true zero rates: on average, their model predicted 173.9 zeroes with 174 actual observed zeroes in the 2009 ACS data. The authors use average MSE to compare their model to the standard SAIPE model, and find that their
model has higher root average MSE for each county population grouping (population grouping range from under 10,000 to over 250,000). Overall, they do not seem to be satisfied with their model and think that it requires future work before it would be ready to be used for official estimates.

A 2012 paper written with Ciara Nugent attempts to address some of the shortcomings of the zero-inflated beta model by adding a random effect to the linking model, but is not entirely successful (Wieczorek, Nugent, and Hawala 2012). In the more recent paper, $\mu_i$ has a distribution around the regression means instead of simply being related by a logit transformation, with an informative prior based on previous years' data. It is difficult to incorporate these features into a beta regression model, causing problems with coding and with MCMC convergence. Again, these models all require that a direct estimate is available for each small area of interest.

Several papers extend hierarchical Bayes small area models to include a spatial or temporal component, such as Kang, Liu, and Cressie (2008). Law et al. (2014) apply a Bayesian spatio-temporal model to crime rates in 1,128 census dissemination areas (DAs) in the city of York, Ontario. DAs are very small areas, with an average population of only 791 persons. Crime data were taken from police reports to the York Regional Police Department for 2006 and 2007, and demographic data were taken from Statistics Canada. Let $Y_{it}$ be the crime rate in DA $i$ ($i = 1, 2, ..., 1128$) and year $t$ ($t=1,2$), $n_{it}$ be the population in DA $i$ for year $t$, with the additional constraint that $n_{t1} = n_{t2}$ (i.e., the population in a
DA is constant from year to year). Then the crime rate for DA \( i \) in year \( t \) can be modeled as

\[
Y_{it} \sim \text{Binomial}(n_{it}, p_{it})
\]

\[
\text{logit}(p_{it}) = \alpha + \beta X_i + u_i + s_i + \gamma t + \delta_i t
\]  

(3.13)

where \( \alpha \) is the mean spatial effect, \( u_i \) are the spatially unstructured random effects for DA \( i \), \( s_i \) are the spatially structured random effects for DA \( i \), \( \gamma \) describes the mean linear time trend, and \( \delta_i \) describes the spatio-temporal trend due to the interaction of DA and time. A simple linear time trend is appropriate here because only two years of data are used; for \( t > 2 \), a more complicated time trend function is probably necessary. The covariate \( X_i \) represents economic deprivation in DA \( i \) as measured by percentage of residents with low family income, and is considered a non-time varying fixed effect.

The model was fit in WinBUGS using noninformative priors on the non-spatially related parameters, and priors determined via an intrinsic conditional autoregressive Gaussian distribution (ICAR) for the spatially related parameters \( s_i \) and \( \delta_i \). Under ICAR, the means of \( s_i \) and \( \delta_i \) are allowed to depend on the means of \( s_i \) and \( \delta_i \) in neighboring DAs and additional variance parameters are determined conditional on the variances of \( s_i \) and \( \delta_i \) respectively in neighboring areas. “Neighbors” were defined as areas sharing one or more common vertex between boundaries.

Law et al. found several areas for which the area-time interaction term \( \delta_i \) was statistically significant, meaning that there was area-specific variability in year to year crime trends. Spatial models make sense at the DA level—these areas are small enough that crime can
easily spill over from one area to the next, and their borders are largely administrative only. The city of York is also covered by a single police agency, so it is probably safe to assume that law enforcement policy is roughly the same across DAs. Spatial modeling is not as useful at the county level; counties are much larger units, and ordinances and laws may differ between counties. Law enforcement agencies often do not cross county borders, so law enforcement policies may also change—a sheriff in one county may routinely require arrests for minor offenses, while the sheriff in the neighboring county may tend to let minor offenders off with a warning. Neighboring counties can also have radically different demographics. For example, Philadelphia County in Pennsylvania coincides with the borders of the city of Philadelphia, a densely populated urban county with a relatively high crime rate. Neighboring Montgomery County is home to some of the wealthiest suburbs of Philadelphia, and is decidedly suburban with family-friendly older neighborhoods, good school districts, and relatively low crime; immediately to the northeast of Philadelphia, Bucks County is known for its farmland, historical sites, and national parks. While the spatio-temporal components of their model may not be directly applicable to county-level estimation, the binomial hierarchical Bayesian model used lends itself well to modeling crime.

Section 3.2: Exploratory analyses and early work
This section contains a summary of exploratory analyses and models that are considered, but that we ultimately decided not to pursue further. In early stages of our research, we attempted to model county-level aggravated assault rates directly, as is done in much of
the previous research summarized in Section 3.1. In Chapters 4 and 5, we present alternative approaches based on estimating the percentage of aggravated assaults reported to the police.

The NCVS collects a great deal of information about each victim and each incident, so we needed to select the best explanatory variables to use to predict county-level crime rates. Age, race, gender, education level, population density, population size, poverty, and transience seemed to be some of the most commonly used variables in similar crime research. We ran logistic regression models on the NCVS data to see which factors seemed to be predictive of aggravated assault victimization in our data file. The response variable is AA, coded as 1 if the person reported at least one aggravated assault victimization in 2011, 0 if not. (Both the NIBRS and NCVS data frames are usually used at the victimization level—so a person reporting two aggravated assaults in 2011 would have two separate records. We reduced the file to a person-level frame for this analysis only so that we could use person-level characteristics.) A separate model was run for persons from the Midwest only, since some later analyses focus on the Midwest.

With only 246 reported aggravated assaults in the 2011 NCVS, it was necessary to collapse some of the categorical variables. Race is divided into white and non-white, age into 12-29, 30-55, and 55+, and education level into less than high school, high school diploma, some college, college degree, or advanced degree. Place size was aggregated into small (0-49,999), medium (50,000-499,999), or large (500,000 or more). The NCVS
variable for household income is missing for nearly half of households, so we did not use income, but investigated “Number of times you have moved in the past five years” as a potential measure of transience. This variable was also missing for most households and was not statistically significant in any models, so it was dropped from the final model.

We used the svyglm function in R from the survey package, which applies the appropriate person-level survey weights to logistic regression (Lumley 2014). The results are in Table 9 below. The results are similar across all cases and the Midwest-only cases for most variables. The reference category for Age is 12-29, so the probability of reporting an aggravated assault victimization decreases with age; white people are less likely to be assaulted than non-white people (although in the Midwest, this is not statistically significant); and since the reference category for MSA status is “Central city,” people living in the central city are more likely to be assaulted than people who live elsewhere.

It seems that once the other variables in the model are accounted for, place size is not an important predictor of aggravated assault. The gender and education estimates seem to disagree between the Midwest-only and full samples. Gender is not statistically significant in the full sample, but it is highly significant in the Midwest-only sample. In the Midwest, all higher education levels (less than high school diploma is the reference level) have a higher probability of victimization for men, while in the full sample men with associate or bachelor's degrees or higher have a reduced probability of victimization.
<table>
<thead>
<tr>
<th></th>
<th>Full sample, n=251463</th>
<th></th>
<th>Midwest only, n=58253</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Std. Error</td>
<td>Pr(&gt;</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>-4.488</td>
<td>0.200</td>
<td>0.00</td>
</tr>
<tr>
<td>Age 30-55</td>
<td>-0.477</td>
<td>0.089</td>
<td>0.00</td>
</tr>
<tr>
<td>Age 55+</td>
<td>-1.436</td>
<td>0.129</td>
<td>0.00</td>
</tr>
<tr>
<td>High school diploma</td>
<td>0.044</td>
<td>0.150</td>
<td>0.77</td>
</tr>
<tr>
<td>Some college</td>
<td>0.026</td>
<td>0.164</td>
<td>0.87</td>
</tr>
<tr>
<td>Associate or bachelor's degree</td>
<td>-0.368</td>
<td>0.166</td>
<td>0.03</td>
</tr>
<tr>
<td>Advanced degree</td>
<td>-1.248</td>
<td>0.378</td>
<td>0.00</td>
</tr>
<tr>
<td>Female</td>
<td>0.095</td>
<td>0.143</td>
<td>0.51</td>
</tr>
<tr>
<td>White</td>
<td>-0.169</td>
<td>0.095</td>
<td>0.08</td>
</tr>
<tr>
<td>Medium place size</td>
<td>0.038</td>
<td>0.122</td>
<td>0.76</td>
</tr>
<tr>
<td>Large place size</td>
<td>0.107</td>
<td>0.159</td>
<td>0.50</td>
</tr>
<tr>
<td>Region: Midwest</td>
<td>0.191</td>
<td>0.139</td>
<td>0.17</td>
</tr>
<tr>
<td>Region: South</td>
<td>0.053</td>
<td>0.126</td>
<td>0.67</td>
</tr>
<tr>
<td>Region: West</td>
<td>0.448</td>
<td>0.133</td>
<td>0.00</td>
</tr>
<tr>
<td>MSA status: (S)MSA but not central city</td>
<td>-0.242</td>
<td>0.119</td>
<td>0.04</td>
</tr>
<tr>
<td>MSA status: Not (S)MSA</td>
<td>-0.213</td>
<td>0.158</td>
<td>0.18</td>
</tr>
<tr>
<td>High school diploma*Female</td>
<td>-0.136</td>
<td>0.205</td>
<td>0.51</td>
</tr>
<tr>
<td>Some college*Female</td>
<td>-0.247</td>
<td>0.227</td>
<td>0.28</td>
</tr>
<tr>
<td>Associate/ bachelor's degree*Female</td>
<td>-0.746</td>
<td>0.247</td>
<td>0.00</td>
</tr>
<tr>
<td>Advanced degree*Female</td>
<td>0.783</td>
<td>0.452</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Table 9: Results from weighted logistic regression predicting aggravated assault victimization in the 2011 NCVS
Also, the interaction terms for the Midwest mean that women are less likely than men to report an aggravated assault, except among those who have less than a high school diploma. This sample does include children 12-18 who have not graduated high school; considering adults without a high school diploma and children still in school separately may be an interesting future analysis.

Aggravated assault victimizations seem to depend most heavily on age, gender, education, and MSA status. With MSA in the model, place size does not have a significant effect. Race did not appear to be statistically significant with all the other variables in the model, but race is enormously important in crime research, so it appears in future analyses. Education level is excluded from most future analyses; although it appears to be important in these models, education data is missing for many NCVS records, and very few people in the sample have less than a high school diploma or an advanced degree.

These preliminary models make the assumption that the NCVS is the “gold standard” for crime data. Many of the papers in the literature on combining two sources of information (for example, Lohr and Brick 2011, Elliott and Davis 2005) assume that one source can be considered the “gold standard”—not exactly the truth, but as accurate as possible. The NCVS captures both reported and unreported crime, so if the NCVS was a large enough survey that reliable county-level estimates were available, we would simply use the NCVS data and perhaps only use NIBRS data to verify. Making this “gold standard”
assumption means that we will use NCVS data to estimate quantities whenever possible, and primarily use NIBRS data to relate crime rates to the county level.

We assume that victimization rates and police reporting rates vary geographically by county, but these rates also depend on covariates like race, age, and sex. An area-level model in which all data are aggregated up to the county level rather than used at the individual level is necessary because of the limitations of the NIBRS data. Let $A_j, j=1, 2, \ldots, J$ be the large areas of interest, defined by a combination of MSA status (non-MSA, in MSA but not central city, central city of MSA), place size (small, under 50,000; medium, 50,000-500,000; large, 500,000 or more), and region (Northeast, Midwest, West, and South). These geographic divisions were chosen because they were the finest level of detail available in the NCVS, and the place size categories were aggregated to give a reasonable number of NCVS victimizations in each cell.

This cross-classification resulted in 21 cells with at least one aggravated assault out of the 36 possible cells; many of the zeros are structural, meaning that zero counties in the United States fell in those cells in 2011. Structural zeros are identified with bold in Table 10, and the shaded cells denote cells with 7 or fewer counties. In the Midwest-only analyses, the large areas will be each combination of MSA status and place size. Nested within each large area $A_j$ are counties $i= 1, 2, \ldots, n_j$ which fall within that MSA status/place size grouping.
<table>
<thead>
<tr>
<th></th>
<th>Northeast</th>
<th>Midwest</th>
<th>South</th>
<th>West</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central city of an (S)MSA</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>In (S)MSA but not in central city</td>
<td>10</td>
<td>12</td>
<td>38</td>
<td>19</td>
</tr>
<tr>
<td>Not (S)MSA</td>
<td>2</td>
<td>12</td>
<td>22</td>
<td>11</td>
</tr>
<tr>
<td>Medium</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central city of an (S)MSA</td>
<td>10</td>
<td>18</td>
<td>19</td>
<td>16</td>
</tr>
<tr>
<td>In (S)MSA but not in central city</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Not (S)MSA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Large</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central city of an (S)MSA</td>
<td>5</td>
<td>12</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>In (S)MSA but not in central city</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Not (S)MSA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 10: NCVS 2011 aggravated assaults, by large area category

It was also necessary to think carefully about our choice of outcome variable. We are interested in modeling the true mean county level aggravated assault rate. If the direct estimates from NIBRS are used as the model outcome, then the quantity estimated is only the mean *reported* aggravated assault rate. For simplicity, we adjusted the NIBRS county level aggravated assault rate upward by dividing by the NCVS percentage of crimes reported to the police in the corresponding large area, resulting in an estimate of the overall county-level aggravated assault rate including crimes not reported to the police. This is admittedly a rough adjustment, but it allowed us to perform some preliminary modeling.
Figure 3: Aggravated assault police reporting rate for geographic cells, 2011 NCVS

The weighted police reporting rates from the NCVS for each large area are plotted in Figure 3, where a reporting rate of 1.0 means that all persons in that cell who reported an aggravated assault also said they reported the assault to the police; a rate of 0.0 means that at least one person reported being assaulted, but no one in that cell reported the assault to the police. Rates very close to 0 or 1 typically denote a cell with very few cases; because some of the weights in the NCVS are so large, any of the weighted rates
could be based on very few cases. There is actually one cell (small, Northeast, not MSA) that contains only two aggravated assault victimizations but has a weighted reporting rate of 0.95; this means that the victimization weight for the police-reported victimization was very large compared to the victimization weight for the non-reported victimization.

The overall weighted police reporting rate for aggravated assaults is about 66%; this disagrees slightly with the published numbers in *Criminal Victimization 2011*, which is expected because BJS works with collection-year estimates while our file is in data-year format. Reporting rates are lower for central cities, medium-sized places, and in the West, and higher in the Northeast, non-MSA area, and small or large places. The cell reporting rates, however, are very volatile because of the small number of cases in some cells. Rather than use the raw estimated rates, we smoothed them towards the respective MSA rate by taking a weighted average of the cell rate and the overall MSA rate based on the standard deviation of each. We chose to smooth towards MSA status because the pattern of police reporting in MSA status is intuitive; it is not as clear why medium-sized places would have a lower police reporting rate than large places, or why the rate for the West is so low. Plots of the raw vs. smoothed rates are in Figure 4, with the overall average rate plotted as a straight line.
We then used the reporting rate for each MSA by size by region cell to adjust the NIBRS crime rate of each county in that cell:

\[
y_{j(i)}^* = y_{j(i)} \ast \left(\frac{1}{b_j}\right)
\]  

(3.14)

where \(y_{j(i)}\) is the NIBRS aggravated assault rate for county \(i\) in MSA by size by region cell \(j\), and \(b_j\) is the NCVS police reporting rate for cell \(j\). A multiplicative adjustment is used rather than an additive one because as the crime rate goes up, it makes more sense that the same proportion of crimes will go unreported than that the same number of crimes will go unreported. Also, under the conditions of Lohr and Brick’s simulation study, models assuming multiplicative bias were consistently robust under true additive bias, but the reverse was not true.
We first tested a zero-inflated beta regression model, adapted from Wieczorek and Hawala's 2011 paper. The model was fitted using the gamlss function in the GAMLSS package (Generalized Additive Models for Location, Scale, and Shape); the link functions used were the default choices in the software package, and the function documentation and relevant examples suggested there was no compelling reason to change them for this analysis (Stasinopoulos et al. 2015). Recall that our goal with this model is to estimate the true (unknown) mean aggravated assault rate in county $i$ nested in large area $j$; call this quantity $\mu_{j(i)}$.

\[
\begin{align*}
y_{j(i)}^* | \mu_{j(i)}, \gamma_{j(i)} &= \begin{cases} 
0 \text{ w.p. } p_{j(i)} \\
\text{Beta}(\mu_{j(i)}, \gamma_{j(i)}) \text{ w.p. } 1 - p_{j(i)}
\end{cases} \\
\text{logit}(\mu_{j(i)}) &= x_i' \beta \\
\text{log}(y_{j(i)}^*) &= a_0 + \alpha_1 (\text{NIBRS covered population}) \\
\text{logit}(p_{j(i)}) &= v_0 + v_1 (\text{NIBRS covered population})
\end{align*}
\]  

In Model 1 described by equation 3.15 above, $y_{j(i)}^*$ is the adjusted NIBRS county-level aggravated assault rate for county $i$ nested in large area $j$, scaled to be 1:100 persons so that the rate would lie between 0 and 1 for all cases (the rates reported for aggravated assault are typically between 0.2 and 0.8 per 100); the beta distribution is parameterized so that $\mu_{j(i)}$ is the true mean county-level crime rate and $\gamma_{j(i)}$ is a parameter related to the variance, as in Wieczorek and Hawala's model; and the covariates are the NCVS aggravated assault rate for large area $j$ along with demographic variables such as percent
white and percent male drawn from the 2011 ACS five-year county-level estimates. The log of the variance parameter depends on the size of the NIBRS-covered population of the county, with the rationale that larger counties would tend to have less yearly variation in crime rates because small counties' rates could be easily influenced by, for example, one big bar fight.

The zero-inflated model is necessary because it is possible for a county to report 0 aggravated assaults and thus a police-reported aggravated assault rate of 0, but the beta distribution only has support on the interval (0, 1). In the 2011 data, 108 counties reported 0 aggravated assaults through NIBRS. The probability of a zero was allowed to depend on the size of the NIBRS-covered population of the county because smaller counties are more likely to report zero assaults. The initial set of covariates for the mean term was selected based on the results of the logistic regression predicting victimization, and variable selection was performed using AIC.

After adjusting the NIBRS crime rates as described in Equation 3.15, the results for fitting Model 1 both to all available counties and to the Midwest only are in Table 11. Most of the estimates agree between all counties and the Midwest counties only. Although the coefficients for NCVS large area aggravated assault rate appear large, the NCVS large area aggravated assault rate is in terms of rate per person, ranging from only about 0.001 to 0.01. For the Midwest-only model, that means that for every 0.001-point increase, we would expect (holding all other variables constant) that the logit of the
adjusted county aggravated rate should increase by about 0.4 points on average. In the middle of the range of rates, this is an increase of only about 0.02-0.06 points in the estimated county rate.

<table>
<thead>
<tr>
<th>All counties (n=1304)</th>
<th></th>
<th></th>
<th></th>
<th>Midwest only (n=551)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Std Error</td>
<td>p value</td>
<td>Estimate</td>
<td>Std Error</td>
<td>p value</td>
<td>---</td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.010</td>
<td>0.5750</td>
<td>0.0792</td>
<td>-4.3783</td>
<td>1.2306</td>
<td>&lt;0.001</td>
<td>---</td>
</tr>
<tr>
<td>NCVS large area AA rate</td>
<td>44.075</td>
<td>10.1760</td>
<td>&lt;0.001</td>
<td>40.2693</td>
<td>12.4620</td>
<td>0.001</td>
<td>---</td>
</tr>
<tr>
<td>Percent white</td>
<td>-1.518</td>
<td>0.1377</td>
<td>&lt;0.001</td>
<td>-0.9913</td>
<td>0.3244</td>
<td>0.002</td>
<td>---</td>
</tr>
<tr>
<td>Percent male</td>
<td>-4.079</td>
<td>1.0724</td>
<td>0.001</td>
<td>1.3378</td>
<td>2.2986</td>
<td>0.561</td>
<td>---</td>
</tr>
<tr>
<td>Percent age 10-29</td>
<td>1.404</td>
<td>0.5323</td>
<td>0.008</td>
<td>1.7464</td>
<td>0.7706</td>
<td>0.024</td>
<td>---</td>
</tr>
<tr>
<td>Intercept, variance model</td>
<td>4.007</td>
<td>0.0402</td>
<td>&lt;0.001</td>
<td>4.405</td>
<td>0.0620</td>
<td>&lt;0.001</td>
<td>---</td>
</tr>
<tr>
<td>Covered pop, variance model</td>
<td>-1.988x10^-7</td>
<td>2.708x10^-7</td>
<td>0.4571</td>
<td>-9.102x10^-7</td>
<td>3.920x10^-7</td>
<td>0.021</td>
<td>---</td>
</tr>
<tr>
<td>Intercept, zero-inflated model</td>
<td>-0.06111</td>
<td>0.0193</td>
<td>0.7514</td>
<td>-0.1257</td>
<td>0.2457</td>
<td>0.609</td>
<td>---</td>
</tr>
<tr>
<td>Covered pop, zero-inflated model</td>
<td>-0.000238</td>
<td>2.817x10^-5</td>
<td>&lt;0.001</td>
<td>-0.000195</td>
<td>3.504x10^-5</td>
<td>&lt;0.001</td>
<td>---</td>
</tr>
</tbody>
</table>

Table 11: Parameter estimates for Zero-Inflated Beta regression, all counties and Midwest only
County-level crime rates also increase on average when a larger percentage of the population is young and non-white. Interestingly, the model for all counties indicates that a larger male population would tend to cause aggravated assault rates to decrease, which contradicts every descriptive statistic in both the NCVS and NIBRS data; we consider this evidence that using all counties probably isn’t appropriate because of the coverage problems in the other regions. The model also indicates that variance decreases as population size increases, and that the probability of observing a zero rate decreases as population size increases, both expected results.

The maps in Figure 5 focus on the state of Ohio, because most (86 out of 88) counties in Ohio have NIBRS data available. (The two counties which have no data are shown in white.) The maps display the raw NIBRS county rates, taken directly from the county-level NIBRS data; the adjusted NIBRS county rates, attempting to account for unreported assaults; and the estimated county-level rates from the Midwest-only model fitted above. See Appendix B for a map of Ohio with county names. The raw rate map has the lowest rates, which is not surprising; the adjusted rates map shows higher rates in the southwest corner of the state (Cincinnati), and in the area of Cleveland, Akron, and Youngstown, as well as a slightly elevated rate in Franklin County. The orange spot in the eastern middle of the state is Muskingum County, which contains the city of Zanesville; the deep orange-red county in the center of the Lake Erie shoreline west of Cleveland is Sandusky County, home of Cedar Point and several lakefront towns and islands known for their summer nightlife. At least for the state of Ohio, the adjusted crime rate map makes sense.
The fitted rate map shows even higher estimated rates of aggravated assault, and identifies urban counties as ones with higher rates. It does pick up Lucas County in the northwest corner of the state—the other two maps put Lucas County in the lowest category, but the city of Toledo in Lucas County should have a relatively high crime rate.
Athens County, in the southeast corner of the state, also shows a relatively high aggravated assault rate—likely because it is the home of Ohio University, a relatively large public university. It would be interesting to be able to compare these results to the NCVS estimate of Ohio's aggravated assault rate; currently, BJS does not publish state estimates. However, weighting the fitted values for the Midwest by county population gives an estimated aggravated assault rate of 0.024 per 10 persons, or 2.4 per 1000 persons—for comparison, the NCVS estimates a rate of 3.9 per 1000 persons for the entire Midwest. This model is still probably underestimating the true rate of aggravated assault.

In fitting Model 1, we thought about each county being divided into several different demographic groups, such as by age, sex, and race—for example, white males 12-29, or black females 55+. The problem then becomes estimating a set of many unknown rates: rather than estimating one county-level rate, we estimate 12 crime rates per county for each of these demographic subgroups, then combine these rates as a weighted average based on county population demographics to get one county-level rate. This is a good theoretical idea because it is known that both crime rates and police reporting rates vary across these demographic groups. In practice, breaking the NCVS data down by only MSA status, size, race (white/nonwhite), and gender results in a sparse table with 20 out of 36 cells containing 0 victimizations. The model is either overly simplified if we drop even more covariates, or there is too little information in each cell to reliably estimate a weighted average rate for each county if more covariates are added.
Model 2 introduces area-specific covariates like MSA and region directly, but incorporates demographic information through a weighted average. We created demographic cells by age category (10-29, 30-54, and 55+), gender (male/female), and race (white/nonwhite). Let k = 1, 2, ..., 12 index these demographic cells, and let i be the index for county i nested in large area j (region by size by MSA). Then calculate:

\[ \varphi_j(i) = \sum_k n_{ik} * r_{jk} \]  

(3.16)

where \( n_{ik} \) is the number of persons in county i in demographic cell k based on the 2011 5-year ACS estimates, and \( r_{jk} \) is the NCVS aggravated assault rate in large area j for demographic group k. \( \varphi_j(i) \) can then be used as an NCVS-based estimate of the crime rate in county i. We use this value as a covariate in place of a raw NCVS crime rate. Rather than using beta regression, we take the logit of the rates and use a standard linear model. The logit of the adjusted NIBRS rates is approximately normal, as seen in the histogram below.

Parameter estimates from the fitted Model 2 for the Midwest are in Table 12; the full sample contained too many counties with missing NIBRS data for reliable estimates. The main effect of the weighted average term is not significantly different from zero, but its interaction with MSA status is—the average effect of being a non-MSA county as compared to a central city MSA county on the logit aggravated assault rate is estimated at 4.5610 – 2.3705* \( \varphi \). Since \( \varphi \) only ranges from 0.002 to 0.04, this model predicts that
non-MSA counties should actually have higher aggravated assault rates than central city counties if all other variables in the model are equal, and that effect decreases slightly as the NCVS weighted average rate increases. However, in practice all other variables are likely not equal between a non-MSA and MSA central city county; if nothing else, percent white is much higher for most rural Midwestern counties than it is for most Midwestern cities.

![Histogram of logit of adjusted NIBRS rates](image)

Figure 6: Histogram of logit of adjusted NIBRS rates

The directions of all the other effects are as expected, although only percent white has a coefficient significantly different from zero. We suspect that entering in the weighted average, MSA status, and the same demographic variables used to construct the weighted average may be causing these variables to interact in strange ways. Removing the demographic variables entirely, however, leads to a poor model fit. The plot of the fitted values (transformed back to the original scale: crime rate per 10 persons) vs. the adjusted
NIBRS rates shows a big clump at low aggravated assault rates. The fitted values are slightly larger than the adjusted rates but the difference isn't as large as in Model 1, and there are still some extreme outliers. The model does not seem to fit particularly well based on the diagnostic plots either (Figure 7). The distribution of the residuals is left-skewed, and the normal probability plot shows some problems with normality.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std Error</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-4.645</td>
<td>3.3636</td>
<td>0.1644</td>
</tr>
<tr>
<td>Phi (NCVS weighted avg)</td>
<td>0.9098</td>
<td>1.1567</td>
<td>0.4319</td>
</tr>
<tr>
<td>MSA but not in central city</td>
<td>-0.5534</td>
<td>2.3896</td>
<td>0.8170</td>
</tr>
<tr>
<td>Not MSA</td>
<td>4.5610</td>
<td>1.0183</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Percent age 10-29</td>
<td>1.5315</td>
<td>2.7124</td>
<td>0.5726</td>
</tr>
<tr>
<td>Percent male</td>
<td>2.5032</td>
<td>3.1313</td>
<td>0.4244</td>
</tr>
<tr>
<td>Percent white</td>
<td>-3.4355</td>
<td>1.7762</td>
<td>0.0537</td>
</tr>
<tr>
<td>Phi*MSA not in city</td>
<td>0.2639</td>
<td>1.1713</td>
<td>0.08218</td>
</tr>
<tr>
<td>Phi*Not MSA</td>
<td>-2.3705</td>
<td>0.4863</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

MSE: 0.9321 on 484 df. Multiple R-squared: 0.1584, Adjusted R-squared: 0.1445

Table 12: Parameter estimates from linear model with response on logit scale, Midwestern counties only

A plot of the estimated aggravated assault rates under Model 2 for Ohio counties, similar to those shown for Model 1, is shown in Figure 8. This model seems to consider MSA status as the driving factor behind aggravated assault rates: the darkest counties are all the counties containing the large Ohio cities. Compared to the map of adjusted rates, this model did not predict higher aggravated assault rates in Sandusky County and identified
Logan County instead, and also tends to minimize the aggravated assault rates in Cleveland's outlying counties while increasing them in Franklin County. Overall, this model does not increase the aggravated assault rates as much as the beta model does, but without knowing the true crime rates it is difficult to say which model is making better predictions.

However, the aggravated assault rate for the entire Midwest based on the NCVS data is 0.0039, or 3.9 per 1000 persons, but the average of the fitted values is only 1.35 per 1000
persons. When weighted by county population, the average of the fitted values is still only 2.6 per 1000 persons—far below the rate predicted by the NCVS. This indicates that this model is still underestimating the true county-level crime rates.

Figure 8: Map of Ohio county-level crime rates, from NIBRS and using Model 2

The models presented in this section are interesting as exploratory analyses on the relationship between NIBRS and NCVS data, but none are promising enough to pursue
further. In Chapter 4 we will present a simulation method that seems to provide an improvement over the results of these models and links NCVS data with county-level NIBRS data in a simple and intuitive way. Chapter 5 introduces a model-based method that we also believe is more successful than any models mentioned thus far.
Chapter 4: A simulation-based method for estimating county-level crime rates

Section 4.1: Introduction to the method and description of underlying process
Recall that the goal of this research is to develop a method for estimating the true county-level crime rate, meaning that both police-reported and unreported crimes are included. The naïve linear regression and beta regression methods used in Section 3.2 do not seem to describe the county-level police reporting rate well. We want to develop a method that will allow us to use data from as many counties as possible while still linking the NCVS data to counties in a meaningful way.

The previous methods all assume that the county-level crime rate is entirely unknown and must be modeled directly. However, the police-reported crime rate at the county level is known when NIBRS data is available; we can reduce the problem to estimating only the rate or number of crimes not reported to the police at the county level. If we develop a reasonable estimate of the percentage of crimes reported to the police, then we can use that estimate along with the NIBRS (police-reported) crime rate to estimate the overall crime rate. In this chapter, we employ a simulation procedure using the NCVS data to estimate the percentage of crimes reported to the police for each county (the county-level police-reporting rate), which is an extension of a technique used by Calder et al. (2008) to estimate mean particulate matter exposure at the county level. We can then estimate the
total county-level number of crimes (reported and unreported) and use ACS estimates of county population to calculate an estimated county-level crime rate.

We restrict our analysis in this chapter to only the crime of aggravated assault for the reasons described in more detail in Section 2.4; they are reviewed briefly here. The definition of aggravated assault is very similar in both the NCVS and NIBRS, and it occurs relatively frequently compared to other types of violent crime but is still serious enough that it is more typically reported to the police. It is helpful to choose a crime with a relatively high reporting rate for model building so that the final estimates depend more heavily on the reported crime rates (assumed known from NIBRS) than on the percentage of crimes reported to the police, which must be estimated. Choosing a crime with relatively robust data that is defined similarly across both data sources also allows us to use published NCVS estimates to check the accuracy of our method. Once we are confident in our method, it can be extended to crimes like simple assault, which tends to be reported at a much lower rate than aggravated assault (in 2011, the NCVS estimated that 67% of aggravated assaults were reported to the police, compared with only 43% of simple assaults), or property crimes, which require an extra step of estimation to make the household-level NCVS data comparable with the person-level NIBRS data.

A simple and intuitive model for number of aggravated assaults reported to the police is of the form

$$Z_i \sim Binomial(\rho_i, Y_i)$$

(4.1)
where $Y_i$ is the true number of crimes that occurred in county $i$ in a given year; $Z_i$ is the total number of crimes reported to the police in county $i$; and $\rho_i$ is the police-reporting rate for county $i$. Our strategy is to use the NCVS data to estimate $\rho_i$, and in turn estimate $Y_i$ for all counties where $Z_i$ is known. Law et al. (2014) also use a binomial distribution to model crime counts for small districts in the city of York, Ontario, but take crime rate and district population as parameters. Their model can be written

$$Z_{it} \sim \text{Binomial}(r_{it}, N_{it})$$

where $N_{it}$ is the population of district $i$ in year $t$ and $r_{it}$ is the crime rate in district $i$, year $t$, accounting for police-reported crimes only.

We chose to model police reporting rate rather than model aggravated assault rate directly for two main reasons. First, aggravated assault rates are (thankfully) low; *Criminal Victimization 2011* estimates that the U.S. aggravated assault rate based on the NCVS was 4.1 per 1000, or a rate of 0.0041. Rates this close to zero can be extremely difficult to model. Police reporting rates tend to be much higher; in 2011, BJS estimated that 68% of aggravated assaults were reported to the police. Second, NIBRS crime counts give us a hard lower bound for the total number of crimes committed in a county. If according to NIBRS there were 100 aggravated assaults reported to the police in county A in 2011, the true number of aggravated assaults in that county should be at least 100, since NIBRS crimes are attributed to the county in which the crime occurred. A binomial model based on county population does not naturally incorporate that lower bound, but a
binomial model based on total number of crimes does; to have observed $Z_i$ successes, there must be $Z_i$ or more total trials.

The difficulty in using the binomial model is that $Y_i$ is obviously unknown and the true police-reporting rate $\rho_i$ is also unknown. However, it is reasonable to assume that $Z_i$ is known from NIBRS. We make the further assumption that the only uncertainty in $Z_i$ for counties fully covered by NIBRS is due to measurement or reporting errors (for example, an aggravated assault that is misclassified as a simple assault, or an agency neglecting to submit NIBRS data for one month). In this situation a negative binomial distribution seems more appropriate. Let $Z_i$ be the number of “successes,” or the number of crimes successfully reported to the police. Then we are interested in the distribution of the total number of all crimes given the number of reported crimes and the police reporting rate for that crime type, $Y_i|\rho_i, Z_i$, which is the total number of trials required to observe $Z_i$ successes given a success rate of $\rho_i$. This distribution is not quite a negative binomial—

$\text{Negative Binomial}(\rho_i, Z_i)$ is defined as the distribution of the number of Bernoulli trials with fixed success rate $\rho_i$ needed to attain $Z_i$ successes, provided that the final trial is a success. We are interested in the distribution of the total number of trials containing $s$ successes, regardless of whether the final trial is a success. This distribution turns out to be the distribution of the total number of trials before the $Z_i+1^{\text{st}}$ success:

---

7 The negative binomial can be parameterized in one of two ways—either as the distribution of the number of failures before observing $y$ successes, or as the distribution of the total number of trials before observing $y$ successes. If $X \sim \text{Negative Binomial}(p, y)$ under the first parameterization and $W \sim \text{Negative Binomial}(p, y)$ under the second parameterization, then $X + y$ is equivalent to $W$. We use the second (total number of trials) parameterization when discussing the negative binomial.
\( Y_i \sim \text{Negative Binomial}(\rho_i, Z_i + 1). \) \hspace{1cm} (4.2)

(See Appendix A for the full derivation). Now estimating \( \rho_i \) for counties with known \( Z_i \) will allow us to make inference about the distribution of \( Y_i \).

For this model to make sense, \( Z_i \) must be a subset of \( Y_i \)—specifically, \( Z_i \) should be the reported crime portion of \( Y_i \). NIBRS and NCVS data are not directly comparable in that sense, as discussed in more detail in Chapter 2. The NCVS excludes crimes committed against children under 12 and businesses, while NIBRS counts such incidents. Fortunately, NIBRS provides information on whether the victim was an individual or a business and records the ages of individual victims, so we can exclude NIBRS crimes against businesses or victims under age 12 from the analysis. We also exclude victims of unknown age from our analysis. This is a benefit of using NIBRS rather than the UCR. Using UCR summary data would allow us to use more counties in our analysis, but would require another step of estimation to exclude crimes committed against victims under 12 or non-individual victims. (See Addington 2007 for more information.)

NIBRS also attributes crimes to the county in which they occurred, while the geographic information available in the NCVS is for the respondent’s place of residence, which is not necessarily where the crime occurred. For example, if a respondent was assaulted in a major city outside of his or her home county, NIBRS would assign the assault to the county containing the major city, while the geographic information in the NCVS record would provide information about the respondent’s home county. In 2011, 44.5% of
NCVS respondents reporting an aggravated assault indicated that the assault took place at or near their place of residence; another 10.6% answered that the assault took place near a friend, neighbor, or relative’s home, and 33.9% said that it occurred in a public area, but there is no way to tell whether these “public areas” are in the respondent’s county of residence or not. We make the simplifying assumption that this source of error is negligible and can be ignored, but should be considered in future work.

We limit our analysis to only counties that contain at least one agency that reports through NIBRS and calculate the police-reported aggravated assault rate as the number of reported aggravated assaults divided by the total population covered by NIBRS. In counties with less than 100% NIBRS coverage, we assume that the police reported aggravated assault rate in the uncovered portions of the county is the same as in the covered portion. This is a reasonable assumption in most cases, since most counties either contain no NIBRS agencies or have over 95% coverage. This results in 1,452 counties with usable NIBRS data.

We also exclude crimes reported to NIBRS agencies that cannot be easily assigned to a county; for example, state police, state highway patrol, or university police forces. We acknowledge that this will cause us to underestimate aggravated assault rates, especially in rural counties that rely on state police as the main law enforcement agency; however, the impact should be slight. There are only 946 aggravated assaults in the 2011 NIBRS file that were reported through agencies missing county-level information, accounting for
less than 0.5% of the 186,287 total aggravated assaults reported through NIBRS in 2011 against individual victims over 12 years old. Maltz (1999) strongly advises against allocating such crimes by population. He uses the example of the state of Connecticut, in which the rural counties are patrolled almost exclusively by the state police, while the large cities have dedicated police forces and rarely need assistance from state troopers. If state police crimes were allocated by population, most crimes would be erroneously allocated to the densely populated cities—but it is also true that the more populous rural counties tend to report more crimes than less populous rural counties, so it is not correct to allocate based on inverse population either. We therefore leave further refinement of this part of our model to those with more subject area expertise. These assumptions are discussed in greater detail in Chapter 2.

Section 4.2: Simulation theory, setup, and exploratory analyses

Without access to the NCVS county-level identifiers, we cannot directly estimate \( \rho_i \) or build a model for \( \rho_i \) based on counties with an NCVS sample. Any direct estimates of \( \rho_i \) would likely be unstable in any case, since there are only 246 total aggravated assaults in the NCVS 2011 data-year file. We also cannot directly apply most small-area estimation methods since NCVS and NIBRS data are not at the same level—the NCVS data is available at the individual level with no useful way to aggregate it to the county level,

---

8 There are an additional 10,060 aggravated assaults reported to agencies that appear to be missing county-level information in the 2011 NIBRS data. However, these agencies are “City” agencies that can be easily assigned to a geographic area. For example, New York City, Baltimore, and St. Louis are all considered independent cities that are not within any county; the state of Virginia also designates 38 independent cities.
and the NIBRS data can only be used at the county level, since it includes no records for individuals who were not crime victims. Instead, we adapt a method proposed by Calder et al. (2008) for linking individual records with county-level data in cases where the county-level identifiers on the individual records are either missing or not useful. Much of the method described by Calder et al. is highly Bayesian and includes a spatial analysis component; we only describe a small portion of the method that is adapted here.

Calder et al. (2008) modeled the effect of particulate matter (PM) concentrations on mortality while controlling for individual PM exposure, rather than simply using ambient outdoor PM concentration like most other models linking PM concentration to mortality. The authors used personal activity records from the National Human Activity Pattern Survey (NHAPS) to estimate individual PM exposure due to sources such as cooking or smoking, as well as to estimate the amount of time an individual spent outside exposed to the ambient PM concentration. Daily county-level air pollution data for eight counties in North Carolina were available so that PM exposure due to time spent outdoors could be calculated, as well as daily county-level covariates such as average temperature and wind speed. The NHAPS data, however, did not contain enough sample cases within the counties of interest so that inference could be made based on those cases alone. The authors instead used 2000 Census data to generate a three-way table of age by sex by employment status for each of the eight counties in the study region, then sampled 100 NHAPS records per county according to the frequency counts in each cell such that the sampled records would have the same age by sex by employment status distribution as
the county population. Previous research had shown that these factors significantly influence activity patterns (and consequently PM exposure), so this sampling method ensured that the activity patterns of the sampled individuals should be reasonably similar to the true activity patterns of county inhabitants. The authors therefore considered the sample of 100 NHAPS records as a proxy sample of individuals from the county of interest. The mean individual PM exposure of the sample could then be considered a county-level PM exposure parameter in higher levels of the model and linked to the other county-level covariates.

We use a similar strategy: generate a random sample of NCVS records for each county based on ACS population data, and let the estimated county-level police-reporting rate \( \hat{\rho}_t \) be the police-reporting rate for aggravated assaults in the sample. Calder et al. determined that a sample of 100 records per county was sufficient to adequately estimate average PM exposure. Aggravated assault, however, is a much rarer event than cooking or being around a smoker; out of the 143,120 personal interviews in the 2011 NCVS, there are only 246 aggravated assaults, and the estimated national rate of aggravated assault in 2011 was only 4.1 per 1000 (Truman and Planty, 2012). Sampling too few NCVS records per county risks drawing a sample that contains no aggravated assaults at all, let alone a sample that contains both police-reported and unreported aggravated assaults. The county-level sample must be large enough so that we can calculate a stable police-reporting rate.
We found that sampling 50,000 NCVS records per county with replacement resulted in roughly 100 aggravated assaults per sample, enough to estimate a relatively stable police-reporting rate. However, sampling too many records from the same year will likely produce biased estimates if the same records are sampled multiple times. This is of special concern for groups typically underrepresented in the NCVS sample, like young non-white males. To avoid this problem, we pooled four years of NCVS interviews from the 2009, 2010, 2011, and 2012 collection-year files to create a larger population available for sampling. By pooling these data we assume that county-level police-reporting rates are relatively stable from year to year across a four-year period. We believe this is a reasonable assumption; although reporting rates may change over time, the change should be gradual enough that the four-year rolling average will not be too different from the rates of the individual years. The combined file was created by stacking the person-level files (DS003) from each of the four years. The final four-year file contained 675,791 interviews, with at least one aggravated assault reported in 861 of the interviews.

By using the person-level files rather than the incident-level files, we include all completed NCVS interviews over the four-year period. It is important that we sample from all NCVS records, not just from the records of aggravated assault victims; exploratory analyses on the NCVS data show that the demographics of crime victims are significantly different from the demographics of all persons in the NCVS sample. Table 13 clearly shows that the total NCVS sample from 2009 through 2012 is largely white.
(82.31%) and includes slightly more interviews from females than males, but only 74.22% of aggravated assault victims are white and 56.91% are male. Since we do not have any demographic information on crime victims at the county-level, only demographic information about the population of the entire county, we need to sample from the entire NCVS population as well rather than only from the population of assault victims.

<table>
<thead>
<tr>
<th>All NCVS interviews, 2009-2012 (N=675,791)</th>
<th>White</th>
<th>Non-White</th>
<th>Total</th>
<th>White</th>
<th>Non-White</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>269222</td>
<td>54879</td>
<td>324101</td>
<td>39.84%</td>
<td>8.12%</td>
<td>47.96%</td>
</tr>
<tr>
<td>Female</td>
<td>287017</td>
<td>64673</td>
<td>351690</td>
<td>42.47%</td>
<td>9.57%</td>
<td>52.04%</td>
</tr>
<tr>
<td>Total</td>
<td>556239</td>
<td>119552</td>
<td>675791</td>
<td>82.31%</td>
<td>17.69%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interviews from aggravated assault victims only (reported at least one AA) (N=861)</th>
<th>White</th>
<th>Non-White</th>
<th>Total</th>
<th>White</th>
<th>Non-White</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>374</td>
<td>116</td>
<td>490</td>
<td>43.44%</td>
<td>13.47%</td>
<td>56.91%</td>
</tr>
<tr>
<td>Female</td>
<td>265</td>
<td>106</td>
<td>371</td>
<td>30.78%</td>
<td>12.31%</td>
<td>43.09%</td>
</tr>
<tr>
<td>Total</td>
<td>639</td>
<td>222</td>
<td>861</td>
<td>74.22%</td>
<td>25.78%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Table 13: Distribution of NCVS interviews by sex and race of respondent, for all respondents and respondents reporting an aggravated assault only, unweighted

Note also that we are sampling interviews, not persons: since an individual in the NCVS sample is interviewed every six months over 3.5 years, a person could have up to seven interview records. We chose not to aggregate the NCVS data to the person level for several reasons. First, if the data are aggregated to the person level, we would need to
somehow correct for the fact that persons are in the sample for different periods of time. If all other characteristics are constant, persons who are in the NCVS sample for two interviews are twice as likely to be victimized while in sample compared to persons who are only in sample for one interview, simply because the time period of interest is twice as long. Demographic characteristics like race, sex, age, and income are often correlated with mobility, so failing to correct for time in sample would cause victimization status to be confounded with these covariates at the person level.

The most straightforward way to account for time in sample is to weight cases proportional to the number of months that the person is in sample. For example, a person who is in sample for 36 out of the 48 months of interest would have a weight proportional to 36/48; a person who is only interviewed once would be in sample for six months and have a weight proportional to 6/48. (This is equivalent to the adjustment recommended in the NCVS codebook when converting NCVS collection-year files to a data-year format; see p. 554 of Bureau of Justice Statistics, 2014, for more detail.) Leaving the records at the interview level preserves this relationship without the need for additional weighting—a person with six interviews is still six times more likely to have an interview selected than a person with only one interview.

Next, police-reporting rate is fundamentally an incident-level parameter rather than a person-level parameter. We want to know what percentage of crime is reported to the police, not what percentage of persons report crimes to the police. It is not possible to
sample at the incident level since non-victimized persons by definition did not report an incident, and as stated before it is important to sample from the population of all persons rather than just the population of crime victims. Sampling at the interview level is as close as we can get to sampling at the incident level while including non-victimized persons, and non-victimized persons must be included if we are matching records based on ACS estimates of total county population. Finally, using interview-level data allows flexibility to sample by variables that may change between interviews. In our application, race, sex, and MSA status do not change between interviews for nearly all persons (recall that the housing unit, not the household, is sampled, so MSA status of a housing unit is by definition unchanged while in the NCVS sample). If we wanted to look at police-reporting rates for domestic violence, however, we might want to sample based on marital status, which could change between interviews. Age and income are other potentially useful covariates that may change while a person is in the NCVS sample. Sampling at the interview level allows us to avoid the messy question of how to properly aggregate such records to the person level.

Without county-level information available for NCVS records, we want to select a sample from the four-year file that is as similar as possible to the true county population in terms of reporting aggravated assaults to the police. For our analysis, we want to select our sample based on factors that significantly influence police-reporting rate. Exploratory analyses on the 2011 NCVS data showed considerable differences in reporting rates based on the victim’s sex and race, as well as the metropolitan statistical area (MSA)
status of the county; see Section 3.2 for more detail. This is consistent with previous research on factors influencing police-reporting rates (Bosick et al. 2012, or Hart and Rennison 2003, for example). We therefore selected our sample based on victim's sex, race, and MSA status of residence; choosing 50,000 NCVS records per county in proportions that agree with American Community Survey (ACS) county population statistics should then result in a sample with a police-reporting rate that is as similar as possible to the true county-level police-reporting rate. (We found that using more than 50,000 records did not improve the simulation results, but did extend the simulation run time.) For example, if the ACS statistics report that county A is 40% white male, 45% white female, 10% non-white male, and 5% non-white female, we randomly select
0.40*50,000=20,000 NCVS records where the respondent was a white male,
0.45*50,000=22,500 NCVS records where the respondent was a white female, and so on. We believe this will produce a good approximation to sampling from the county population itself in the absence of actual county-level identifiers.

Proportions of county population by sex and race used for sampling were based on ACS 5-year estimates of county population 10 years of age and older by sex and race. MSA status was assigned to counties based on the Census Bureau’s 2012 classification file. Children under 10 were excluded from the ACS data to make the sampling proportions as close as possible to the proportions in the NCVS population of all individuals 12 years of age and older; ACS tables were only available with five-year age groupings, and any difference in sex by race proportions caused by including ten and eleven year olds should
be negligible. To ensure that there were enough cases of aggravated assault in all cells, race was further collapsed into white/non-white and MSA status into central city (CC) or non-central city (non-CC), where non-central city includes both non-central city counties within an MSA and non-MSA counties. Once the sample of NCVS cases was taken, the estimated police-reporting rate $\hat{\rho}_i$ among aggravated assaults in county $i$ was calculated as

$$\hat{\rho}_i = \frac{\sum_{j=1}^{50000} AA_j^r}{\sum_{j=1}^{50000} AA_j}$$

where $j$ runs over the 50,000 NCVS records sampled for county $i$; $AA_j^r$ is the number of police-reported aggravated assaults for record $j$; and $AA_j$ is the total number of aggravated assaults for record $j$. In other words, we take the total number of police-reported aggravated assaults in the sample and divide it by the total number of aggravated assaults in the sample. Note that while rare, it is possible for a record to include more than one aggravated assault; a respondent assaulted twice in the six-month reference period for that record may have reported both, neither, or only one of the assaults to the police. In that case we would count both aggravated assaults in the denominator, but would count only the police-reported assault(s) in the numerator.

Table 14 provides the number of interviews available for sampling in each sex by race by MSA status cell, along with the number of interviews with at least one aggravated assault reported by cell. All sampling is performed with replacement. The number of available
interviews with non-white respondents appears low; if we are sampling 50,000 interviews per county and each county is classified as either Central City or non-Central City, then there is some concern that counties where a large proportion of the population is non-white will be sampling from the same small pool of records. This would cause the estimates for these counties to be more similar than they should otherwise be. We do not, however, believe this will be a problem in our analysis. Of the 1,452 counties included in the analysis, no county had a non-white male population of more than 42% of the total population 10 and over or a non-white female population of more than 40%, so in the most extreme scenario we sample 21,000 out of 25,667 non-white male Central City interviews and 20,000 out of 30,697 non-white female Central City interviews. The median non-white male population is 3.8% of total population, and the corresponding median non-white female population is 2.9%, so for the vast majority of counties we sample 25% or fewer of the available interviews in a given cell.

Table 14 again demonstrates the importance of sampling from all NCVS interviews, rather than only from interviews including one or more aggravated assaults. The distribution of interviews with assault victims is quite different from the distribution of all interviews—for example, 30.77% of all interviews were with white females living in non-Central City counties, but they make up only 19.40% of interviews with assault victims.
<table>
<thead>
<tr>
<th></th>
<th>All interviews (N=675,791)</th>
<th>Interviews including at least one aggravated assault (N=861)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Central City</td>
<td>Non-Central City</td>
</tr>
<tr>
<td><strong>Counts</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>White</td>
<td>73,574</td>
</tr>
<tr>
<td></td>
<td>Non-white</td>
<td>25,667</td>
</tr>
<tr>
<td>Female</td>
<td>White</td>
<td>79,106</td>
</tr>
<tr>
<td></td>
<td>Non-white</td>
<td>30,697</td>
</tr>
<tr>
<td><strong>Percentages</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>White</td>
<td>10.89%</td>
</tr>
<tr>
<td></td>
<td>Non-white</td>
<td>3.80%</td>
</tr>
<tr>
<td>Female</td>
<td>White</td>
<td>11.71%</td>
</tr>
<tr>
<td></td>
<td>Non-white</td>
<td>4.54%</td>
</tr>
</tbody>
</table>

Table 14: Available NCVS Interviews and Interviews with Aggravated Assaults

Section 4.3: Simulation output and adjustments

To summarize, we performed the following procedure for each of the 1,452 counties in our analysis:

- Classify the county as Central City or non-Central City, based on the Census Bureau designation;
- Use ACS 5-year demographic tables to determine what percentage of the county population was white male, white female, non-white male, and non-white female;
- Based on these percentages, sample 50,000 interviews with replacement from a file of all 675,791 NCVS interviews conducted from 2009-2012. For example, if...
the county of interest was a Central City county, with ACS estimates of 40% white male, 40% white female, 10% non-white male, and 10% non-white female, we would sample 20,000 NCVS interviews from white males living in Central City counties, 20,000 NCVS interviews from white females living in Central City counties, and 5,000 NCVS interviews each from non-white males and females living in Central City counties;

- Obtain the total weighted number of police-reported aggravated assaults in the sample of 50,000 interviews (AA'), and the total weighted number of all aggravated assaults (AA) in the sample. The estimated weighted police-reporting rate for the county was then calculated by dividing AA' by AA.

We were concerned about the variability of the simulated police-reporting rate generated by this process, so the simulation was run five times to check the stability of the estimated rates, resulting in a set of five simulated reporting rates per county. The average standard deviation of the estimated police-reporting rate across the five runs was 5.13%, meaning that the average margin of error on each police-reporting rate is 10%--an unacceptably large interval. To minimize dependence on one particular sample realization and reduce the overall variability of our estimates, we decided to use the mean police-reporting rate across the five simulation runs for each county. Essentially, we are taking the mean of five draws from the sampling distribution for each county, giving us a more accurate estimate of the true mean police-reporting rate. The estimated mean county-level police-reporting rates for nearly all counties changed very little after three
simulation runs, so the estimated mean rates can be considered stable at five runs. Williamson, Birkin and Rees (1998) likewise used a five-run average when comparing sample optimization methods; see Section 3.1 for more detail.

Figure 9 shows histograms of the simulated police-reporting rates over all 1,452 counties in the sample. The left-hand plot is a histogram of the rates from the first trial only; the corresponding histograms from the other four trials were very similar, so they are omitted. The histogram on the right shows the average simulated police-reporting rates across five trials. Variability is reduced by taking the average rate, but the distribution of the average rates is now clearly bimodal.

Upon investigation, we found that the two peaks can be explained by looking at Central City (CC) counties and non-Central City (non-CC) counties separately—the distribution of rates for all counties is really a mixture of the distributions for CC and non-CC counties. The same bimodal pattern occurs for single trials, but the greater variability within a single trial causes the peaks to overlap in such a way that the distribution appears unimodal. The peak in the histogram at lower police reporting rates corresponds to non-CC counties, while the simulated rates for CC counties are somewhat higher—a counterintuitive result, since we expected that people in urban areas would be generally less likely to report crimes to the police than those in suburban or rural areas.
A preliminary investigation showed that this issue may have been caused by using the raw, unweighted NCVS data for sampling. The NCVS has a complex, multi-stage cluster design with weights that must be used when analyzing the NCVS data alone. Initially, we had hoped that our sampling strategy would allow us to use the unweighted data, but the NCVS weights incorporate important design information beyond broad race category, sex, and MSA status. We, therefore, decided that it was important to incorporate the sampling weights in our analysis.

Weighted estimates were generated by applying the victimization weight to all aggravated assaults. The weighted police-reporting rate is given by:
\[
\hat{\rho}_{i(Weighted)} = \frac{\sum_{j=1}^{50000} \sum_{k=0}^{AA_j} w_{jk}^r}{\sum_{j=1}^{50000} \sum_{k=0}^{AA_j} w_{jk}}
\]  

(4.4)

The subscript \(j\) indexes records in the sample and runs from 1 to 50,000; the subscript \(k\) indexes aggravated assaults within record \(j\) and runs from 0 to \(AA_j\), the total number of aggravated assaults reported for record \(j\). In our data, \(k\) runs from 0 to 5 since the minimum number of aggravated assaults per record is 0 and the maximum is 5. Let \(w_{jk}\) be the victimization weight associated with aggravated assault \(k\) in record \(j\), and define \(w_{j0}\) as 0 so that records with no aggravated assaults do not contribute to the sum. Let \(w_{jk}^r\) be the victimization weights associated with police-reported aggravated assaults only, where \(w_{j0}^r = 0\) and \(w_{jk}^r = 0\) if the \(k^{th}\) aggravated assault was not reported to the police, \(w_{jk}^r = w_{jk}\) otherwise. Then the weighted police-reporting rate for county \(i\) is the sum of the weights of all police-reported aggravated assaults in the sample drawn for county \(i\) divided by the sum of the weights of all aggravated assaults in the sample.

Figure 10 shows the simulated police-reporting rates under both unweighted and weighted conditions for all available U.S. counties. We used the same set of simulated data to calculate both rates, meaning that use of the weights is the only difference between the two sets of rates. In each case, we find that non-CC counties tend to have lower simulated reporting rates than CC counties. We found that recent research using NCVS data supports this finding. Bosick et al. (2012) found that when controlling for other variables like victim age, gender, race, and type of offense, violence was generally
more likely to be reported by victims living in urban areas than in suburban or rural areas. Baumer and Lauritsen (2010) found a similar result when looking at police-reporting trends in the NCVS from 1973 to 2010.

The weighted police-reporting rates are somewhat lower than the unweighted rates, both overall and by MSA status. This is likely because the NCVS weights are designed in part to adjust for underrepresented populations, such as the young or the transient, and such populations are generally less likely to report crime to the police. Using the weights also incorporates information from the NCVS's complex, multistage cluster design; all
published NCVS estimates use weighted victimization data. Because of this, we consider the weighted rates to be more reliable and appropriate in this application.

Finally, we performed a nonparametric bootstrap procedure to obtain a rough estimate of the variability of the mean weighted county-level police reporting rates. For each county, we resampled the five simulation estimates of police-reporting rate with replacement and calculated the mean police-reporting rate of this resample. We performed this procedure 1000 times per county, resulting in 1000 bootstrap samples and therefore 1000 mean police-reporting rates calculated for each county. These 1000 means for county \( i \), \( \hat{\rho}_i^k \), \( k = 1, 2, \ldots 1000 \), were ordered and used to construct nonparametric 90% and 95% confidence intervals for the simulated police-reporting rate for each county. To estimate confidence intervals for the mean population-weighted aggravated assault rates for each county \( i \) we drew from \( \text{Negative Binomial}(\hat{\rho}_i^k, Z_i + 1) \) for each \( k=1, 2, \ldots 1000 \), then again ordered these draws to construct nonparametric 90% and 95% confidence intervals.

It is not practical to report confidence intervals for each county in our sample, but we are able to use the bootstrap samples to calculate nonparametric confidence intervals for the population-weighted means. For example, the confidence intervals for the population-weighted mean for Ohio were calculated by restricting the dataset to only the 86 counties in the state of Ohio and finding the population-weighted mean for each of the 1000 sets of bootstrap samples. These 1000 means were ordered as before, and the 95% confidence interval taken as the 26th ordered mean to the 975th, so that the middle 950/1000 means
are covered. This is admittedly a rough estimate of the variability of our procedure, and further work on the properties of our method is necessary.

Section 4.4: Results

With an estimate of the aggravated assault police reporting rate for each county, \( \hat{\rho}_i \), we can now make use of Equation 4.2. Assume that the true number of aggravated assaults in county \( i \), \( Y_i \), has a negative binomial distribution with parameters \( Z_i + 1 \) and \( \hat{\rho}_i \). We use the mean of this negative binomial as a simple estimate of the total number of aggravated assaults in county \( i \), \( \hat{Y}_i \). Naively dividing the number of police-reported aggravated assaults by the estimated reporting rate gives a simple estimate of total aggravated assaults, but results in an estimate of exactly zero total aggravated assaults for any county with zero aggravated assaults reported through NIBRS. Using the mean of the negative binomial is still a straightforward calculation, but avoids this problem for counties with zero NIBRS counts. The mean is calculated as:

\[
\hat{Y}_i = (Z_i + 1) \ast \left( 1 + \frac{1 - \hat{\rho}_i}{\hat{\rho}_i} \right).
\] (4.5)

We can now apply this step to our simulation results to estimate the total number of aggravated assaults (reported and unreported) for counties in our sample. All analyses above were performed using all available data (n=1,452 counties), but we describe the simulation results for three geographic areas: the state of Ohio, the Midwest region, and the entire United States. We first present results from the state of Ohio, as we consider
those results to be the most complete and reliable. Out of 88 counties, 86 Ohio counties had NIBRS reporting rates of 90% or higher, so Ohio’s NIBRS crime rates can be considered fairly complete and reliable. Ohio also is one of the few states where the largest cities (Columbus, Cleveland, and Cincinnati) all report through NIBRS, so we can clearly see differences between Central City and non-Central City counties. We also present results from the Midwest and from all available counties in the U.S. (all counties with at least one NIBRS reporting agency); however, these results should be considered much less reliable than those for Ohio because of the NIBRS coverage problems mentioned before. Recall that large agencies like New York City, Chicago, Los Angeles, Detroit, and most other major U.S. cities do not report through NIBRS. Since aggravated assault rates tend to be higher in large metropolitan areas and their large populations give them considerable influence on the population-weighted rates, we would expect our estimated average rates to be somewhat lower when such agencies are excluded. County-level maps are not presented for either the Midwest or the U.S. since coverage is so sparse.

The first column in Table 15 shows summary statistics for NIBRS aggravated assault rates from all 86 available Ohio counties. The NIBRS aggravated assault rates for Ohio tend to be lower than the national NIBRS aggravated assault rates; the national population-weighted mean aggravated assault rate was 1.95 per 1000 in 2011 (given in

---

9 Washington D.C. appears to report some NIBRS data; however, the only reporting agency is the Metro Transit Police, which is responsible for only D.C.-area transit facilities and does not have any associated covered population.
Table 16), while the corresponding population-weighted rate for Ohio was only 1.09 per 1000. This in turn will cause the estimated aggravated assault rates for Ohio to be lower than the national estimates, or the estimates in *Criminal Victimization 2011*.

<table>
<thead>
<tr>
<th></th>
<th>Raw NIBRS</th>
<th>Simulated police reporting rates (weighted)</th>
<th>Estimated agg assault rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0.0000</td>
<td>44.31%</td>
<td>0.0180</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>0.1647</td>
<td>53.41%</td>
<td>0.3350</td>
</tr>
<tr>
<td>Median</td>
<td>0.3428</td>
<td>57.84%</td>
<td>0.6210</td>
</tr>
<tr>
<td>Mean</td>
<td>0.5185</td>
<td>57.30%</td>
<td>0.9257</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>0.5902</td>
<td>60.59%</td>
<td>1.0910</td>
</tr>
<tr>
<td>Max</td>
<td>2.4750</td>
<td>67.05%</td>
<td>4.4170</td>
</tr>
<tr>
<td>Population weighted mean for Ohio</td>
<td>1.0916</td>
<td>59.19%</td>
<td>1.8578</td>
</tr>
<tr>
<td>95% confidence interval</td>
<td>-</td>
<td>(53.71%, 65.19%)</td>
<td>(1.6342, 2.1321)</td>
</tr>
<tr>
<td>90% confidence interval</td>
<td>-</td>
<td>(54.44%, 64.19%)</td>
<td>(1.6647, 2.0833)</td>
</tr>
</tbody>
</table>

Table 15: Distribution of NIBRS-only and estimated aggravated assault rates per 1000 people, reporting rates in percentages. Ohio only (n=86 counties).

We will primarily refer to population-weighted means for comparison, since weighting by county population provides a much more accurate picture of the true crime rate over the entire state than simply taking the average crime rate over all counties. The sample contains a large number of rural, low-population counties with few aggravated assaults; a simple mean gives too much weight to these counties, while underestimating the influence of large counties with typically higher aggravated assault rates. We present
summary statistics for both the unweighted and weighted simulated police-reporting rates for comparison, as well as the corresponding weighted and unweighted adjusted aggravated assault rates. The distribution of the weighted police-reporting rates for Ohio is slightly less variable than the distribution of the unweighted police-reporting rates, and is centered at a lower rate (59.19% vs. 64.06%). Lower simulated police-reporting rates in turn cause the weighted adjusted aggravated assault rates to be generally higher than the unweighted rates—given the same NIBRS (police-reported) aggravated assault rate, a county with a lower estimated reporting rate will have a higher estimated true (reported and non-reported) rate of aggravated assault. For example, if counties A and B each have a NIBRS (reported) aggravated assault rate of 2.0 per 1000, but in county A only 50% of aggravated assaults are reported to the police, while in county B 60% of aggravated assaults are reported, county A has a higher total (actual) aggravated assault rate (4.0 for A vs. 3.33 for B).

The maps below in Figure 11 show the raw, or unadjusted, NIBRS aggravated assault rates by Ohio county, along with the estimated aggravated assault rates by county. The simulated weighted police-reporting rates are also presented. The crime patterns are as expected; the darker red (higher crime rate) counties generally correspond to the large metro areas in Ohio. All Ohio cities with a population of 80,000 or greater are labeled on the maps. (See Appendix B for a map of Ohio with county names.) Franklin County, where Columbus is located, is in the center of Ohio—the higher aggravated assault rate is not obvious in the raw NIBRS data, but is picked up by our model. Our model also
identifies Muskingum County, east of Franklin County, and a handful of counties in southern Ohio as having higher aggravated assault rates than the rest of the state; these counties have had problems with drug activity, so it makes sense that assaults would rise as well (OCJS 2011).

Figure 11: Maps of Ohio Counties showing NIBRS Aggravated Assault Rates, Estimated Actual Aggravated Assault Rates, and Estimated Police-Reporting Rates
Interestingly, the map of simulated police-reporting rates roughly corresponds to maps of median income or percent in poverty for Ohio—in general, the counties with the highest police-reporting rates tend to be wealthier counties, and the counties with the lowest rates tend to be the poorer counties, even though our method does not explicitly account for income. (See Larrick 2014 for more information on Ohio's county-level poverty rates.) For example, southeastern Ohio is in the Appalachian region, which tends to be more disadvantaged than the rest of the state, and many of the counties with relatively low police-reporting rates are clustered there. Northeastern Ohio and the Cleveland suburbs have some of the wealthiest communities in the state, and also some of the highest police-reporting rates in the state. The county with the highest police-reporting rate for aggravated assault (67.05%) is Ottawa County on Lake Erie, which is a popular summer vacation destination in Ohio with less than 10% of the population in poverty; the county with the lowest weighted police-reporting rate (44.31%) is Vinton County in southeast Ohio, one of the poorest counties in Ohio with over 21% of residents below the poverty level.

This is consistent with a recent BJS report looking at the relationship between poverty and police-reporting rates using NCVS data from 2008-2012; the report found that in rural areas, violent crime was generally less likely to be reported by the poor (Harrell et al. 2014). (However, the same report found that in urban areas the relationship is reversed—persons in high-income households were actually less likely to report violent crime to the police.) We are not able to directly include income in our analysis due to the
amount of missing income data in the NCVS, so it is encouraging to see that the relationship between our estimates and income is in the expected direction.

Expanding the analysis to the Midwestern states, we predict a population weighted mean aggravated assault rate for the Midwest of 3.53 per 1000. Summary statistics for counties with NIBRS data in the Midwest are given in Table 16. The county with the highest aggravated assault rate, both in the NIBRS data and after adjustment, is Jackson County, MO, which is just outside of Kansas City. The high NIBRS rate (10.15 per 1000) is the primary reason for this. The Midwestern counties with the highest reporting rates (over 65%) also tended to have very low NIBRS aggravated assault rates, with the exception of the counties containing Wichita, KS, Nashville, TN, and Milwaukee, WI, which all had simulated reporting rates greater than 65% but NIBRS aggravated assault rates of 4.0 per 1000. For comparison, the U.S. aggravated assault rate in 2010 was estimated to be 3.4 per 1000 by the NCVS, with an estimated 60% of aggravated assaults reported to the police; in 2011 the estimate was 4.1 per 1000 with an estimated 67% reporting rate. Our predicted population-weighted means for the Midwest both for weighted police-reporting rates and for weighted adjusted aggravated assault rate are comparable to NCVS national estimates for 2010 and only slightly lower than the 2011 estimates.

We urge caution in interpreting the estimates for the Midwest, however, due to the number of counties with missing NIBRS data. Out of 1,054 total counties in the Midwest, data were only available for 551. Exploratory analyses in Chapter 2 showed that
missingness for Midwestern agencies was fairly evenly distributed in terms of place size and MSA status, so it may be reasonable to assume missingness at random, but further analysis would be necessary. Maltz and Targonski (2002) illustrate in greater detail how missing agency data can seriously bias any analysis that simply ignores missingness, and we share their reservations on interpreting results with a large amount of missing data.

<table>
<thead>
<tr>
<th></th>
<th>Raw NIBRS</th>
<th>Simulated police reporting rates (weighted)</th>
<th>Estimated agg assault rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0.0000</td>
<td>42.37%</td>
<td>0.0180</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>0.2984</td>
<td>51.23%</td>
<td>0.6531</td>
</tr>
<tr>
<td>Median</td>
<td>0.7853</td>
<td>54.14%</td>
<td>1.6130</td>
</tr>
<tr>
<td>Mean</td>
<td>1.1000</td>
<td>55.04%</td>
<td>2.1030</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>1.5610</td>
<td>59.03%</td>
<td>2.9480</td>
</tr>
<tr>
<td>Max</td>
<td>10.1500</td>
<td>68.38%</td>
<td>18.3100</td>
</tr>
<tr>
<td>Population weighted mean</td>
<td>2.0819</td>
<td>59.01%</td>
<td>3.5579</td>
</tr>
</tbody>
</table>

95% confidence interval: (53.58%, 64.90%)  (3.1328, 4.0953)
90% confidence interval: (54.42%, 63.85%)  (3.1968, 3.9945)

Table 16: Distribution of NIBRS-only and estimated aggravated assault rates per 1000 people, reporting rates in percentages. Midwest only (n=551 counties)

Table 17 shows our results for all available U.S. counties: the 1,438 counties with NIBRS data, out of a total of 3,007 U.S. counties. Again, we cannot consider the missing counties to be missing at random and therefore ignorable, as discussed in more detail in Chapter 2. The Northeast has far more counties with missing NIBRS data; only 15% of agencies in the Northeast report through NIBRS, compared to over 54% of agencies in
the Midwest, 38% in the South, and 32% in the West. Large agencies serving a population of over 100,000 persons are also slightly less likely to report through NIBRS than smaller agencies, with 37% of smaller agencies reporting through NIBRS but only 29% of larger agencies. This is especially concerning because large agencies typically serve major cities with relatively high crime rates, and omitting such agencies will cause our estimates to be systematically too low. Despite these serious limitations, our estimates for population weighted mean reporting rate and adjusted aggravated assault rates are still relatively close to the NCVS 2011 national estimates.

<table>
<thead>
<tr>
<th>Raw NIBRS</th>
<th>Simulated police reporting rates (weighted)</th>
<th>Estimated agg assault rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0.0000</td>
<td>0.0180</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>0.4277</td>
<td>0.9044</td>
</tr>
<tr>
<td>Median</td>
<td>0.9747</td>
<td>1.9120</td>
</tr>
<tr>
<td>Mean</td>
<td>1.4080</td>
<td>2.6500</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>1.8230</td>
<td>3.4150</td>
</tr>
<tr>
<td>Max</td>
<td>16.9300</td>
<td>29.6000</td>
</tr>
<tr>
<td>Population weighted mean for the U.S.</td>
<td>1.9562</td>
<td>3.3179</td>
</tr>
<tr>
<td>95% confidence interval</td>
<td>-</td>
<td>(54.07%, 65.40%)</td>
</tr>
<tr>
<td>90% confidence interval</td>
<td>-</td>
<td>(54.86%, 64.46%)</td>
</tr>
</tbody>
</table>

Table 17: Distribution of NIBRS-only and estimated aggravated assault rates per 1000 people, reporting rates in percentages. All available U.S. counties (n=1438)
In this chapter, we have presented a relatively intuitive and easily generalizable method for estimating county-level crime rates using publicly-available data. The simulation process for estimating police-reporting rates is also not computationally intensive: we were able to simulate 50,000 draws for each of over 3,000 counties in about 6 hours using R’s sample package. (We chose to run the simulation for all U.S. counties rather than only counties with NIBRS sample, but only included NIBRS counties in our estimates.)

The results of our method applied to 2011 NIBRS aggravated assault data agree with expected aggravated assault trends: higher rates in urban areas, lower rates in rural areas. Using four years of NCVS data to estimate police-reporting rates will have a smoothing effect on trends in police-reporting rates, but using yearly NIBRS data means that this method should still pick up year-to-year variation in county-level crime trends.

A weakness of this method is the lack of NIBRS data available. The FBI and BJS, however, are actively trying to recruit more agencies, especially large agencies, to NIBRS since implementation has largely stalled. BJS is sponsoring the National Crime Statistics Exchange (NCS-X), which provides grants and assistance to help law enforcement agencies switch to NIBRS reporting, as well as other incentives like optimizing agency resources during NIBRS implementation. NCS-X was rolled out over summer 2014 in 400 sampled agencies, after which BJS will assess its impact and consider extending the program to other non-NIBRS agencies. BJS has made it clear with NCS-X that they view NIBRS as the future of police-reported crime statistics, so it makes sense to rely on NIBRS to develop methods, with the assumption that NIBRS coverage
will continue to improve over the next several years. The BJS website provides more detailed plans at http://www.bjs.gov/content/ncsx.cfm.

It is also difficult to assess more precisely how accurate our estimates are, since there are currently no county- or state-level gold standards for comparison. Fortunately, BJS is performing a pilot study to improve NCVS state-level estimation by increasing sample in 11 states (including Ohio) to permit direct estimation, and plans to have 22 states with enough NCVS sample cases for direct estimation by 2016 (Planty 2014). Direct NCVS estimates would allow us to check our estimates at the state level, rather than having only national estimates for comparison.

The simulation procedure described in this chapter is attractive because of its simplicity, but it may be too simplistic to properly describe county-level crime rates. It does not take advantage of the hierarchical nature of these data (counties nested in larger geographic areas), and does not provide a comprehensive model for the relationship between county demographics, large-area characteristics, and police reporting rates. We attempt to remedy some of these shortcomings in the next chapter. Chapter 5 presents a formal hierarchical statistical model with a well-specified error structure, which allows us to estimate the error in our simulated values more precisely than a simple bootstrap procedure.
Chapter 5: Model-based approaches for estimating county-level crime rates

Section 5.1: Introduction and justification

The simulation-based method described in Chapter 4 is simple, intuitive, and appears to perform reasonably well, but it is not based on a formal statistical model. A successful model should result in predictions that are reasonably consistent with the simulation-based predictions and with published results in *Criminal Victimization 2011*, given the limitations of the available data. It should also incorporate the hierarchical structure of the data; for example, counties should be nested within the corresponding region. We would like our final model to have parameters that are easily interpretable, including a set of parameters that explain the variability in the model. It must be relatively simple: the available NCVS data on aggravated assaults not reported to the police is very sparse, and complex models fit to sparse data run the risk of overfitting. The final model should also be easily generalizable to data from other years, other covariates, and other crime types.

In this chapter we attempt to build a statistical model to predict county-level aggravated assault rates that does incorporate some of these desirable features. Due to the limitations of our data, it is difficult to find a standard statistical model that fulfils all these criteria; instead, we adapt modeling strategies into estimation procedures in some sections.
We take the same strategy as in Chapter 4: assume that the county-level police reported crime rate is known through NIBRS, and attempt to estimate the county-level police reporting rate, \( \hat{\rho}_{j(i)} \). Preliminary models and descriptive analyses in Section 5.2 lay the foundation for the models in the next three sections of this chapter. Sections 5.3 and 5.4 describe models that were fitted as part of the model building process, but were not entirely successful. We want a model that is as simple as possible, yet accounts for the variability that is inherent in NCVS crime rates. We begin in Section 5.3 with the most straightforward model, but move on to more complex models in Section 5.4 and 5.5 as we find that the basic models do not incorporate enough variability in the final estimates. Descriptions of these models and results are included because they introduce important components of the final procedure. Section 5.5 presents the final procedure, adapted from a hierarchical Bayes model combining a normal prior with the same negative binomial setup used in Chapter 4, which we believe fulfills all the criteria above as well as possible.

Section 5.2: Preliminary estimates of police reporting rate at the large area level and at the county level

All models in this chapter rely on three major components: the number of police-reported crimes, estimates of aggravated assault police reporting rate at the large area (MSA by region by place size) level, and prior estimates of aggravated assault police reporting rate at the county level. The number of police-reported crimes is available from NIBRS, but the other two components must be estimated. As before, let \( A_j, j=1, 2, \ldots, J \) be the large
areas of interest, defined by a combination of MSA status (non-MSA, in MSA but not central city, central city of MSA), place size (small, under 50,000; medium, 50,000-500,000; large, 500,000 or more), and region (Northeast, Midwest, West, and South).

Direct large-area estimates of the aggravated assault police reporting rates, \( \rho_j \), are not reliable for all cells; although we refer to the MSA by region by place size cells as large areas, disaggregating the 246 total aggravated assault victimizations in the 2011 NCVS data into the 36 large area cells results in a relatively sparse table with some zero cells. As explained in Section 3.2, six cells can be considered structural zeroes, since no U.S. counties fall under that combination of place size, MSA status, and region. Structural zeros are identified with bold in Table 18, and the shaded cells denote cells with 7 or fewer counties in the United States, which can be considered nearly structural zeros. (Table 18 is identical to Table 10 in Section 3.2, but repeated here for convenience.)

<table>
<thead>
<tr>
<th></th>
<th>Northeast</th>
<th>Midwest</th>
<th>South</th>
<th>West</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central city of an (S)MSA</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>In (S)MSA but not in central city</td>
<td>10</td>
<td>12</td>
<td>38</td>
<td>19</td>
</tr>
<tr>
<td>Not (S)MSA</td>
<td>2</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Medium</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central city of an (S)MSA</td>
<td>10</td>
<td>18</td>
<td>19</td>
<td>16</td>
</tr>
<tr>
<td>In (S)MSA but not in central city</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Not (S)MSA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Large</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central city of an (S)MSA</td>
<td>5</td>
<td>12</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>In (S)MSA but not in central city</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Not (S)MSA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 18: NCVS 2011 aggravated assaults, by large area category
The NCVS data on aggravated assault victimizations is so sparse that small-area techniques are required even for what we refer to as the large geographic areas. Direct estimates of the sample police reporting rate for aggravated assault for most of the nonzero cells have unacceptably wide confidence intervals; for the 19 out of 21 cells with a sample size of 20 cases or fewer, the width of the 95% confidence interval for the direct estimate is larger than 0.4 (or 40%). This means that a 95% confidence interval based on the direct estimates could range from a 50% reporting rate to a 90% reporting rate. We apply a Fay-Herriot model to reduce the variability of these estimates. We would also like to choose covariates for our Fay-Herriot model that allow us to make predictions for some of the nine large areas that have no observed aggravated assaults (and thus no direct NCVS estimates) but do have counties with NIBRS data.

Model covariates are somewhat limited at the large-area level. The variables used to define large areas—MSA status, place size, and region—are available, but these are all categorical variables. The main effects of these variables alone do not describe police reporting rate particularly well, and with only 21 observations adding interaction terms uses up the limited degrees of freedom without significantly improving model fit. We looked at NCVS aggravated assault police reporting rates over time to determine what other factors could be used to predict large-area aggravated assault police reporting rate. Specifically, we were interested to see if the police reporting rates for aggravated assault were correlated with police reporting rates from other, more frequently occurring crimes. Police reporting rates for aggravated assault and other crimes for the years 1993 through
2013 are available through the BJS NCVS Victimization Analysis Tool (NVAT) Custom Tables feature. It is also possible to extract police reporting rates by MSA status, region, or place size. We compared aggravated assault police reporting rates to the police reporting rates for other individual crime types over that period, as well as to the police reporting rates for violent crimes excluding aggravated assault and for property crimes.

Figure 12 displays trend lines for selected crime types, using published BJS estimates from 1993-2013. The trend lines for simple assault and violent crime excluding aggravated assaults are relatively flat compared to the trend line for aggravated assault. This is likely because simple assaults and violent crimes include far more yearly victimizations than aggravated assault; in 2011, for example, there were 1,117 violent victimizations including 964 simple assaults. The trend line for burglary in Figure 12, however, seems to be slightly more variable and tends to follow the same general pattern as aggravated assault. In fact, of all the combinations of crime types, we found that burglary police reporting rates are the best single predictor of aggravated assault police reporting rates across this time period as seen in Table 19.
This is somewhat surprising because burglary is considered a property crime, while aggravated assault is a personal crime. Burglary is the most serious of all the property crimes, however—it requires either forcible entry or unlawful entry without personal contact. For example, if a person comes home from work to find his computer missing, the crime will be classified as a burglary regardless of whether the burglar broke a window or walked in an unlocked front door. If the same computer was stolen while unattended in a coffee shop, it would be classified as a theft. (And if the person was at home when the burglar broke in and was personally threatened in any way, the crime would be classified as a robbery and a personal crime). Perhaps victims consider burglary and aggravated assault to be comparably severe and report them to the police similarly.
<table>
<thead>
<tr>
<th>Crime type</th>
<th>Total</th>
<th>Urban</th>
<th>Suburban</th>
<th>Rural</th>
<th>North-east</th>
<th>Midwest</th>
<th>South</th>
<th>West</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violent crime excl. AA</td>
<td><strong>0.442</strong></td>
<td>0.106</td>
<td><strong>0.435</strong></td>
<td>0.050</td>
<td>0.173</td>
<td>0.019</td>
<td>0.106</td>
<td>0.056</td>
</tr>
<tr>
<td>Rape</td>
<td>0.145</td>
<td>0.186</td>
<td>0.178</td>
<td>0.050</td>
<td>0.230</td>
<td>0.197</td>
<td>0.282</td>
<td><strong>-0.534</strong></td>
</tr>
<tr>
<td>Robbery</td>
<td>0.341</td>
<td>0.281</td>
<td>0.289</td>
<td>0.091</td>
<td><strong>0.474</strong></td>
<td>-0.149</td>
<td>0.112</td>
<td>-0.368</td>
</tr>
<tr>
<td>Simple assault</td>
<td>0.421</td>
<td>-0.036</td>
<td>0.407</td>
<td>0.062</td>
<td>0.021</td>
<td>-0.082</td>
<td>0.087</td>
<td>0.322</td>
</tr>
<tr>
<td>Personal theft</td>
<td>0.128</td>
<td>0.263</td>
<td>0.071</td>
<td>-0.253</td>
<td>0.250</td>
<td>-0.403</td>
<td>0.203</td>
<td>-0.140</td>
</tr>
<tr>
<td>All property crime</td>
<td><strong>0.462</strong></td>
<td><strong>0.660</strong></td>
<td>0.315</td>
<td>0.003</td>
<td><strong>0.515</strong></td>
<td>0.020</td>
<td>0.331</td>
<td>0.101</td>
</tr>
<tr>
<td>Burglary</td>
<td><strong>0.573</strong></td>
<td><strong>0.643</strong></td>
<td>0.370</td>
<td>0.089</td>
<td>0.293</td>
<td>-0.005</td>
<td>0.147</td>
<td>0.122</td>
</tr>
<tr>
<td>Motor vehicle theft</td>
<td>0.280</td>
<td>0.387</td>
<td>0.023</td>
<td>-0.035</td>
<td>0.360</td>
<td>0.358</td>
<td>0.075</td>
<td><strong>-0.471</strong></td>
</tr>
<tr>
<td>Theft</td>
<td>0.421</td>
<td><strong>0.598</strong></td>
<td>0.271</td>
<td>-0.152</td>
<td><strong>0.476</strong></td>
<td>-0.128</td>
<td>0.398</td>
<td>0.207</td>
</tr>
</tbody>
</table>

Table 19: Pearson correlations between police reporting rate for aggravated assault and police reporting rate for selected crimes, overall and by MSA and by region. **Bold** text indicates the correlation is significant at the 95% level.

We find the same relationship in most cases when looking at police reporting rates by region or by MSA status in Table 19 and in Figure 13. Burglary police reporting rate is reasonably highly correlated with aggravated assault police reporting rate for urban and suburban areas, although it performs less well across regions. We considered property crime and theft as alternatives to burglary since those rates are also reasonably highly correlated with aggravated assault across regions and MSA status, but these crime types do not perform well in models: the trends for property crime and theft police reporting rates are relatively flat over time due to the large number of reported victimizations for each type.
Figure 13: Police reporting rate of selected crime types by region and MSA status, NCVS 1993-2013
There are 1,146 reported burglary victimizations in the 2011 NCVS. By including police reporting rate for burglary in the model, we are able to make predictions for three additional cells which do not have any aggravated assault victimizations reported in 2011. (See Table 20. Cells in bold are structural zeroes, meaning that no U.S. counties fall into that category.) If we fit a model based only on MSA, place size, and region, we would be able to make predictions for all 36 cells, but predictions for cells with 0 aggravated assault victimizations would be unreliable since they would require extrapolation based on broad categorical variables. Including burglary in the model means that only six counties with NIBRS sample fall in the cells excluded from the model (that is, cells that have no aggravated assault victimizations and no burglary victimizations in the 2011 NCVS data file); we consider this evidence that the data available for these cells are too sparse for reliable modeling, and exclude these cells from our analysis.

<table>
<thead>
<tr>
<th></th>
<th>Northeast</th>
<th>Midwest</th>
<th>South</th>
<th>West</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>Central city of an (S)MSA</td>
<td>7</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>In (S)MSA but not in central city</td>
<td>44</td>
<td>96</td>
<td>170</td>
</tr>
<tr>
<td></td>
<td>Not (S)MSA</td>
<td>21</td>
<td>72</td>
<td>111</td>
</tr>
<tr>
<td>Medium</td>
<td>Central city of an (S)MSA</td>
<td>22</td>
<td>71</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>In (S)MSA but not in central city</td>
<td>7</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Not (S)MSA</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Large</td>
<td>Central city of an (S)MSA</td>
<td>15</td>
<td>37</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>In (S)MSA but not in central city</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Not (S)MSA</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 20: NCVS 2011 burglaries by large area category.
Once we selected covariates by looking at trends over time, we fit the Fay-Herriot model using the 2011 NCVS data only. Each observation in this data set consists of a large area (MSA by place size by region) with an associated police reporting rate for aggravated assault, for a total of 21 observations from cells with at least one aggravated assault victimization in the 2011 NCVS. As in the preliminary models in Section 3.2, place size and MSA status are highly correlated and should not both be included in the model. This leaves police reporting rate for burglary, MSA status, region, and all interaction terms as potential explanatory variables for aggravated assault police reporting rate. Stepwise regression in R was then used to select the best linear model using this combination of predictors, with “best” defined as minimizing AIC.

The best linear model selected included the interaction term for MSA by region. However, this interaction term requires six degrees of freedom (3 non-reference levels of region by 2 non-reference levels of MSA status) out of 20 available degrees of freedom. The slight improvement in model fit was not worth the risk of overfitting with such a small sample size, and the interaction term made fitting the more complicated Fay-Herriot model difficult, so we decided to drop the interaction term from the final model.

The direct estimates of police reporting rate were also transformed to the logit scale. Direct estimates for some cells were close to zero or one (for three cells, exactly equal to either zero or one), which resulted in Fay-Herriot estimates that were slightly negative or
greater than one in some cases—inappropriate for proportions. A logit transformation allows modeling on the entire real line, while ensuring that the final estimates will be in the interval (0, 1) when transformed back to the original scale. We chose to exclude two cells from the analysis that had direct estimates for reporting rate of exactly one, since the logit of one is infinity, and one cell with a direct estimate of exactly zero since the R package used to fit the model would not handle zeroes; each of these cells included only one NCVS victimization, so little information was lost. This left 18 cells including 243 reported aggravated assault victimizations used for model fitting. The final area-level Fay-Herriot model used to estimate the large-area police reporting rate $\rho_j$ for aggravated assault can be written as

$$
\text{logit}(\rho_j) | \theta_j \sim N(\theta_j, \psi_j), \quad \theta_j | \beta, \sigma^2 \sim N(X_j \beta, \sigma^2), \quad (5.1)
$$

$$
X_j \beta = \beta_0 + \beta_1 \ast \text{MSA: suburban}_j + \beta_2 \ast \text{MSA: rural}_j + \beta_3 \ast \text{region: NE}_j + \beta_4 \ast \text{region: S}_j + \beta_5 \ast \text{region: W}_j + \beta_6 \ast \text{burglary}_j
$$

In Model 5.1, “MSA:suburban” and “MSA: rural” are indicator variables for MSA status of large area $j$, and likewise the “region:” variables are indicator variables for the three regions, with Midwest as the reference category. Again, the regression coefficients $\beta$ are given an improper uniform prior, and we assume that the $\text{logit}(\rho_j)$ are conditionally independent given $\theta_j$, and the $\theta_j$ are conditionally independent given $\beta, \sigma^2$. This model assumes that the covariates $X_j$ and the sampling variances $\psi_j$ are known; as before, in practice, the sampling variances must be estimated using the victimization replicate
weights contained in the NCVS victim-level file. The variance parameter in the linking model, $\sigma^2$, is estimated from the data. $\theta_j$, the true county-level mean, is the quantity of interest but is not observed directly—only $\rho_j$ is observed.

As explained in Section 3.1, Fay and Herriot’s 1979 paper showed that the James-Stein estimator for $\theta_j$ is given by

$$logit(\tilde{\rho}_{jFH}) = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \psi_j} logit(\rho_j) + \frac{\psi_j}{\hat{\sigma}^2 + \psi_j} logit(\tilde{\rho}_j)$$ (5.17)

where $\hat{\sigma}^2$ is the sample estimate of $\sigma^2$, $\rho_j$ is the direct survey estimate for cell j, and $\tilde{\rho}_j$ is the regression estimate of $\rho_j$—essentially, a weighted average of the sample and regression estimates. The R package sae was used for model fitting (Molina and Marhuenda 2015).

We compared the estimates from four Fay-Herriot models to assess the effect of the logit transformation and the effect of adding burglary police reporting rate as a covariate: two models in which the aggravated assault police reporting rate was modeled directly, one including police reporting rate for burglary as a covariate and one omitting burglary, and the same two models using a logit transformation on the response. The estimates from the rate models tend to be lower than the estimates from the logit models, with a few exceptions, and in general the logit models seem to have more of a smoothing effect than the models using police reporting rate directly, as seen in Figure 14. It is difficult to tell
based on the plot alone which model performs best, so we also calculated the mean squared prediction error (MSPE) for each model, computed as $\sum_i (\rho_i - \tilde{\rho}_i)^2$. Lower values of MSPE indicate that predicted values are closer to the corresponding direct estimates, generally indicating better model fit. Before calculation, the estimates for the logit model were transformed back to the original rate scale so that MSPE would be directly comparable between models. Table 21 also reports AIC and parameter estimates for each model, but these are not comparable across the rate and logit models because of the change in scale.

Figure 14: Fay-Herriot model comparison
MSPE is admittedly a weak measure of fit; under a Fay-Herriot model, predictions that are far from the corresponding direct estimates often occur because the direct estimate is based on very few cases and is thus unreliable. We report MSPE because there are few alternatives, especially with response variables on different scales, but use caution in its interpretation. Visual inspection of the plot and MSPE both indicate that including burglary police reporting rate as a covariate tends to improve the model, although AIC increases slightly in the logit models when burglary is added, so we retain burglary police reporting rate as a covariate. Both measures also indicate that the rate and logit models with burglary included are comparable, so we choose the model using the logit transformation because it also has the desirable property of guaranteeing predictions on the interval $(0, 1)$.

<table>
<thead>
<tr>
<th></th>
<th>Rate model</th>
<th>Rate model, burglary omitted</th>
<th>Logit model</th>
<th>Logit model, burglary omitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.357</td>
<td>0.542</td>
<td>-0.277</td>
<td>1.151</td>
</tr>
<tr>
<td>MSA: Suburban</td>
<td>0.203</td>
<td>0.183</td>
<td>0.152</td>
<td>-0.289</td>
</tr>
<tr>
<td>MSA: Rural</td>
<td>0.321</td>
<td>0.221</td>
<td>0.330</td>
<td>-0.066</td>
</tr>
<tr>
<td>Region: NE</td>
<td>0.232</td>
<td>0.151</td>
<td>0.495</td>
<td>0.534</td>
</tr>
<tr>
<td>Region: South</td>
<td>0.189</td>
<td>0.003</td>
<td>0.070</td>
<td>-0.152</td>
</tr>
<tr>
<td>Region: West</td>
<td>0.118</td>
<td>-0.181</td>
<td>0.027</td>
<td>-0.252</td>
</tr>
<tr>
<td>Burglary reporting rate</td>
<td>1.419</td>
<td>-</td>
<td>1.946</td>
<td>-</td>
</tr>
<tr>
<td>AIC</td>
<td>7.175</td>
<td>7.848</td>
<td>50.257</td>
<td>48.939</td>
</tr>
<tr>
<td>MSPE</td>
<td>0.390</td>
<td>0.483</td>
<td>0.383</td>
<td>0.481</td>
</tr>
</tbody>
</table>

Table 21: Fay-Herriot model comparison
We now take the estimates from this model as $\hat{p}_j$, the estimated police reporting rate for NCVS aggravated assault in large area $j$, and build on them for the next level of modeling. The 18 large areas with usable predictions cover 1,446 out of the 1,451 U.S. counties with NIBRS coverage. Because only five counties fall in the three large areas with direct estimates of exactly zero or one that are excluded, we do not attempt to make predictions for the missing large areas, especially since we would be predicting aggravated assault police reporting rate for combinations of categorical variables (MSA status and region) that we do not observe. Such predictions would be unreliable at best and seriously biased at worst, and it is better to not be able to produce estimates for five counties in our data than to make bad predictions for model building.

Section 5.3: Simple negative binomial model and results

As before, let the number of police-reported crimes in large area $j$ be $Z_j$, and the number of police-reported crimes in county $i$ nested in area $j$ is denoted by $Z_{j(i)}$. Similarly, let $Y_j$ and $Y_{j(i)}$ denote the corresponding total numbers of crimes, whether reported or unreported, and $\rho_j$ and $\rho_{j(i)}$ denote the police reporting rates. The most straightforward way to connect the large-area estimates of police reporting rate with the counties nested in each large area is to make the very strong assumption that all counties within a given large area have the same aggravated assault police reporting rate, which must be the same as the overall police reporting rate for the large area itself. That is,
\[ \rho_j = \rho_{j(i)}, \]

\[ i = 1, 2, 3 \ldots n_j \text{ for all counties } i \text{ in large area } j. \] (5.3)

Then the proposed model for the number of police-reported crimes at the county level,

\[ Z_{j(i)} \sim Bin(Y_{j(i)}, \rho_{j(i)}) , \]

simplifies to \[ Z_{j(i)} \sim Bin(Y_{j(i)}, \rho_j) . \] As explained in Chapter 4, an equivalent model more suited to the problem is \[ Y_{j(i)} \sim Negative \ Binomial(Z_{j(i)} + 1, \hat{\rho}_j) , \]

since the number of police-reported crimes or “successes,” \[ Z_{j(i)} \], is known from NIBRS, while the total number of crimes \[ Y_{j(i)} \] is unknown. Simply plugging in the estimates \( \hat{\rho}_j \) from Section 5.2 allows us to estimate the distribution of \( Z_{j(i)} \).

As in Chapter 4, the mean of \( Negative \ Binomial(Z_{j(i)} + 1, \hat{\rho}_j) \) is taken as the point estimate for the true total number of aggravated assaults (reported and unreported) in county \( i \) nested in large area \( j \), and corresponding county-level estimates for aggravated assault rates are calculated by dividing the estimated number of aggravated assaults by the ACS 5-year estimate of total county population 10 and over. We again look at estimates for Ohio counties only and for Midwestern counties only. Estimates for all available U.S. counties are also presented, but should be interpreted with extreme caution because the counties missing due to lack of NIBRS data cannot be considered missing at random. (See Chapter 2 for a more thorough discussion of this issue.) We repeated all analyses by using random draws from the negative binomial distribution rather than the mean and found that the final estimates changed very little; in most cases, they were consistent with the mean-based estimates down to the hundredths place. This is to be
expected, since this procedure is essentially a Monte Carlo estimate of the mean. Since the estimates are so similar, we present only the mean-based estimates because they are more easily replicable.

Table 22 compares the simulation-based estimates from Chapter 4 with the estimates from this model, restricted to the 86 counties with available NIBRS data in the state of Ohio. The table also includes the relevant raw NIBRS aggravated assault rates for comparison. The coarseness of the simple negative binomial model is obvious from the distribution of police reporting rates: the first quartile and the median are equal. In fact, because all Ohio counties are within the Midwest region, there are only 6 different estimates for police reporting rate used in this model. The point estimates for police reporting rates based on large area status alone are somewhat higher than the simulation-based point estimates of police reporting rate, which in turn causes the point estimates for aggravated assault rates to be lower than in the simulation-based model.

The point estimate for population-weighted mean police reporting rate, which we consider the best metric of overall police reporting rate for the state of Ohio, is thirteen percentage points higher under the negative binomial model (59.19\% for the simulation-based model vs. 72.42\% for the negative binomial model). This leads to nearly a 0.4 point change in estimated population-weighted mean aggravated assault rate (1.86 per 1000 vs. 1.48 per 1000). A 0.4 point per 1000 change in aggravated assault rate may seem small, but applied over Ohio’s 2011 population of approximately 11.5 million it
translates to a difference of 46,000 predicted aggravated assaults. Recall from Chapter 4 that we were concerned that the simulation-based estimates of aggravated assault rate were somewhat low when compared to published BJS statistics, so a model with even lower predicted aggravated assault rates is not desirable.

<table>
<thead>
<tr>
<th></th>
<th>Raw NIBRS agg assault rates</th>
<th>Police reporting rates</th>
<th>Agg assault rates</th>
<th>Police reporting rates</th>
<th>Agg assault rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0.0000</td>
<td>44.31%</td>
<td>0.0180</td>
<td>66.67%</td>
<td>0.0180</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>0.1647</td>
<td>53.41%</td>
<td>0.3350</td>
<td>72.63%</td>
<td>0.2654</td>
</tr>
<tr>
<td>Median</td>
<td>0.3428</td>
<td>57.84%</td>
<td>0.6210</td>
<td>72.63%</td>
<td>0.4982</td>
</tr>
<tr>
<td>Mean</td>
<td>0.5185</td>
<td>57.30%</td>
<td>0.9257</td>
<td>73.63%</td>
<td>0.7402</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>0.5902</td>
<td>60.59%</td>
<td>1.0910</td>
<td>75.61%</td>
<td>0.8273</td>
</tr>
<tr>
<td>Max</td>
<td>2.4750</td>
<td>67.05%</td>
<td>4.4170</td>
<td>77.92%</td>
<td>3.4520</td>
</tr>
<tr>
<td>Population weighted mean, Ohio</td>
<td>1.0916</td>
<td>59.19%</td>
<td>1.8578</td>
<td>72.42%</td>
<td>1.4833</td>
</tr>
<tr>
<td>95% confidence interval</td>
<td>-</td>
<td>(53.71%, 65.19%)</td>
<td>(1.6342, 2.1321)</td>
<td>(49.80%, 90.87%)</td>
<td>(1.2251, 2.4153)</td>
</tr>
<tr>
<td>90% confidence interval</td>
<td>-</td>
<td>(54.44%, 64.19%)</td>
<td>(1.6647, 2.0833)</td>
<td>(54.02%, 89.20%)</td>
<td>(1.2492, 2.1895)</td>
</tr>
</tbody>
</table>

Table 22: Distribution of NIBRS-only, simulation-based estimates, and negative binomial model-based estimates of aggravated assault rates per 1000 people and police reporting rates in percentages. Ohio only (n=86 counties).

Further, the confidence bounds for the point estimates on the negative binomial model are extremely wide compared to the bounds on the simulation-based model. The confidence
bounds were calculated via a method similar to a bootstrap, using the boot package in R. For the reporting rates, each set of simulation runs for a county were sampled with replacement 1000 times, resulting in a set of 1000 bootstrap means per county, which could then be used to calculate a set of 1000 population-weighted mean police reporting rates. The 26th and 975th records from this ordered set were taken as the upper and lower nonparametric 95% confidence bounds. The process was similar for the aggravated assault rates, except with the additional step of drawing from the appropriate negative binomial distribution for each replicate.

The table of Midwestern counties only (Table 23) and the table of all available counties (Table 24) show similar patterns to the Ohio-only table. In each case the negative binomial model predicts much higher police reporting rates and much lower aggravated assault rates than the simulation-based model, and shows little variability in police reporting rate. It seems that the available geographic information alone (place size, MSA, and region) does not adequately account for the county-level variability in aggravated assault police reporting rates. This makes sense; the geographic information is at a very high level, with only 18 cells for 1,446 counties. Our exploratory analyses also show that while police reporting rate for aggravated assault does differ based on geographic variables, it is affected even more strongly by demographic characteristics like sex, age, and race. We need to incorporate more county-level information into the model, as demonstrated in the next section.
<table>
<thead>
<tr>
<th></th>
<th>Raw NIBRS agg assault rates</th>
<th>Police reporting rates</th>
<th>Agg assault rates</th>
<th>Police reporting rates</th>
<th>Agg assault rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0.0000</td>
<td>42.37%</td>
<td>0.0000</td>
<td>66.67%</td>
<td>0.0180</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>0.2984</td>
<td>51.23%</td>
<td>0.5722</td>
<td>74.25%</td>
<td>0.5184</td>
</tr>
<tr>
<td>Median</td>
<td>0.7853</td>
<td>54.14%</td>
<td>1.4780</td>
<td>74.25%</td>
<td>1.1860</td>
</tr>
<tr>
<td>Mean</td>
<td>1.1000</td>
<td>55.04%</td>
<td>1.9740</td>
<td>74.37%</td>
<td>1.6120</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>1.5610</td>
<td>59.03%</td>
<td>2.7680</td>
<td>76.07%</td>
<td>2.2080</td>
</tr>
<tr>
<td>Max</td>
<td>10.1500</td>
<td>68.38%</td>
<td>18.3100</td>
<td>77.92%</td>
<td>14.1500</td>
</tr>
<tr>
<td>Population wtd mean, Midwest</td>
<td>2.0819</td>
<td>59.01%</td>
<td>3.5325</td>
<td>72.72%</td>
<td>2.9585</td>
</tr>
<tr>
<td>95% confidence interval</td>
<td>-</td>
<td>(53.58%, 64.90%)</td>
<td>(3.2153, 3.8952)</td>
<td>(52.48%, 91.52%)</td>
<td>(2.3090, 4.5613)</td>
</tr>
<tr>
<td>90% confidence interval</td>
<td>-</td>
<td>(54.42%, 63.85%)</td>
<td>(3.2699, 3.8300)</td>
<td>(56.53%, 89.97%)</td>
<td>(2.3546, 4.1343)</td>
</tr>
</tbody>
</table>

Table 23: Distribution of NIBRS-only, simulation-based estimates, and negative binomial model-based estimates of aggravated assault rates per 1000 people and police reporting rates in percentages. Midwest only (n=555)

<table>
<thead>
<tr>
<th></th>
<th>Raw NIBRS agg assault rates</th>
<th>Police reporting rates</th>
<th>Agg assault rates</th>
<th>Police reporting rates</th>
<th>Agg assault rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0.0000</td>
<td>41.74%</td>
<td>0.0180</td>
<td>60.28%</td>
<td>0.0180</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>0.4277</td>
<td>51.51%</td>
<td>0.9044</td>
<td>68.78%</td>
<td>0.7239</td>
</tr>
<tr>
<td>Median</td>
<td>0.9747</td>
<td>55.01%</td>
<td>1.9120</td>
<td>72.51%</td>
<td>1.4750</td>
</tr>
<tr>
<td>Mean</td>
<td>1.4080</td>
<td>55.56%</td>
<td>2.6500</td>
<td>72.23%</td>
<td>2.1450</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>1.8230</td>
<td>59.64%</td>
<td>3.4150</td>
<td>74.25%</td>
<td>2.7010</td>
</tr>
<tr>
<td>Max</td>
<td>16.9300</td>
<td>73.38%</td>
<td>29.6000</td>
<td>82.93%</td>
<td>62.8100</td>
</tr>
<tr>
<td>Population wtd mean, all counties</td>
<td>1.9562</td>
<td>59.52%</td>
<td>3.3179</td>
<td>72.28%</td>
<td>3.0246</td>
</tr>
<tr>
<td>95% confidence interval</td>
<td>-</td>
<td>(54.07%, 65.40%)</td>
<td>(2.9113, 3.8696)</td>
<td>(50.42%, 90.47%)</td>
<td>(2.3240, 5.0287)</td>
</tr>
<tr>
<td>90% confidence interval</td>
<td>-</td>
<td>(54.86%, 64.46%)</td>
<td>(2.9683, 3.7730)</td>
<td>(52.27%, 88.84%)</td>
<td>(2.3709, 4.781)</td>
</tr>
</tbody>
</table>

Table 24: Distribution of NIBRS-only, simulation-based estimates, and negative binomial model-based estimates of aggravated assault rates per 1000 people and police reporting rates in percentages. All available U.S. counties (n=1446)
Section 5.4: Beta-negative binomial model and results

We take an empirical Bayesian approach to incorporate county-level information in our model. Assume, as in Section 5.3, that the number of all aggravated assaults for county \( i \) nested in large area \( j \), \( Y_{j(i)} \), is distributed as \( \text{Negative Binomial}(Z_{j(i)} + 1, \rho_{j(i)}) \). Then as before, the number of police-reported aggravated assaults \( Z_{j(i)} \) is considered known from NIBRS and \( \rho_{j(i)} \), the police reporting rate for county \( i \) nested in large area \( j \), is considered unknown and must be estimated. Rather than using only the coarse geographic information to link NCVS data to the county level and simply setting \( \rho_{j(i)} = \rho_j \), we consider another way to connect NCVS data to individual counties. We know from the simulation-based method in Chapter 4 that incorporating demographic information like race and sex improves estimation of police reporting rates; a Bayesian model allows us to integrate this demographic information in a cohesive framework. The general idea is to construct a prior distribution for aggravated assault police reporting rate based on the characteristics of the county population, then update this prior based on large-area information. However, we do not have county-level data to use for the updating; step we only have county-level estimates based on NCVS large-area information. Because of this, we no longer have a Bayesian model in the traditional sense: instead, we propose a procedure based on Bayesian modeling techniques.

The simulation-based method used ACS county-level demographic information to match NCVS records to counties. In this procedure, we use the ACS county-level demographic information together with NCVS data aggregated over the same demographic
characteristics to estimate the mean and standard deviation of the distribution of aggravated assault police reporting rates for each county in our data. Under the simulation-based method in Chapter 4, the simulation step was used to link the NCVS data and the individual counties; taking a simple weighted average of mean NCVS police reporting rates for each demographic cell would have too much of a smoothing effect on the final estimates. The Bayesian procedure in this section allows us to use a simple weighted average because it is used for the prior, then updated with geographic estimates to add variability.

Previous research on police reporting rates, as well as exploratory analyses in Chapter 2, suggest that race, age and sex play an important role in reporting to the police. (See also Section 3.2 for more a more detailed description of why these factors are important in modeling aggravated assault police reporting rate.) Poverty status and transience are other potential factors, but the measures of income, poverty status, and transience in the NCVS are all plagued by missing data. With only 246 total aggravated assault victimizations in the 2011 NCVS data year file, we cannot afford to discard any observations; we also cannot assume that such variables are missing at random, since highly transient persons are less likely to respond, as are persons with very low or very high incomes. Race, age and sex of victim are reported for every NCVS victimization and are likely to be accurately recorded, since there is little reason to lie about one’s race, age, or sex, nor is it likely that a respondent will forget these personal characteristics. As in Chapter 4, we aggregate race into white and non-white only because there are so few NCVS records in
the non-white subcategories. We also aggregate age into three categories: 12-29, 30-54, and 55+; as explained in Section 3.2, exploratory analyses showed that the percentage of aggravated assaults reported to the police varied among these three groups. To avoid estimates based on very small cells we restrict our covariates to race, age, and sex only; more covariates could be added, but caution must be used to ensure that cell sizes are large enough for stable estimates.

Let \( k = 1, 2, \ldots, 12 \) index the cells in the race (white, non-white) by age category (12-29, 30-54, 55+) by sex (male, female) table. Then the weighted mean police reporting rate based on demographic factors for county \( i \) nested in large area \( j \) is constructed as

\[
\hat{\mu}_{j(i)}^{NCVS} = \frac{\sum_k n_{ik} \cdot \hat{r}_k}{\sum_k n_{ik}}
\]

(5.4)

where \( n_{ik} \) is the number of persons in county \( i \) in demographic cell \( k \) based on the 2011 5-year ACS estimates and \( \hat{r}_k \) is the estimated NCVS police reporting rate for aggravated assault for demographic group \( k \). \( \hat{\mu}_{j(i)}^{NCVS} \) can then be used as an NCVS demographically-based estimate of the police reporting rate in county \( i \).

It is also possible to estimate the standard deviation of this mean by using the replicate weights included in the NCVS data files. The NCVS uses a complex, multi-stage survey design, and the exact sampling procedure is not public due to confidentiality concerns. However, clustering and stratification must be taken into account during analysis for
accurate point estimates and confidence intervals. The victimization weights are designed to produce accurate point estimates, and similarly the replicate weights are generated based on the sample design to produce accurate confidence intervals when used with replication methods. The estimated police reporting rate for cell k, \( \hat{r}_k \), is calculated by summing the victimization weights of all aggravated assault victimizations reported to the police and dividing them by the sum of all the victimization weights of all aggravated assaults, reported and unreported. We can repeat this rate calculation for each of the 160 sets of victimization weights to achieve a set of 160 rates, \( \hat{r}_k^n \), n=1, 2, 3,…160. Then we follow the method for using the replicate weights laid out in the NCVS Technical Documentation (2014):

\[
\hat{\sigma}(\hat{r}_k) = \frac{4}{160} \sum_{n=1}^{160} (\hat{r}_k - \hat{r}_k^n)^2 
\]  

(5.5)

The documentation notes that the factor of 4 is necessary because of the way that the replicate weights were constructed; see p. 115 of the NCVS Technical Documentation for more details. We can perform an analogous calculation to estimate the covariances between rates, \( \text{cov}(\hat{r}_k, \hat{r}_l) \) for k=1, 2, …,12, and l=1, 2,…,12, with k\neq l. It is then straightforward to apply rules for the variance of sums to Formula 5.4 and obtain estimates for the sample variance of \( \hat{\mu}_{j(i)}^{NCVS} \) in each county; call these estimates \( \hat{\sigma}_{j(i)}^2 \).

Now that we have estimates for the prior means and variances for each county of interest, we must select an appropriate distribution. The Beta distribution is a conjugate prior for
the negative binomial, and it also has the nice property of having support on \((0, 1)\), so in this procedure we make the simplifying assumption that the police reporting rate for county \(i\) is distributed \textit{a priori} as \( Beta(\alpha_{(i)}, \beta_{(i)}) \). We use the method-of-moments estimators for \(\alpha_{(i)}\) and \(\beta_{(i)}\) based on the estimated county-level sample means, \(\hat{\mu}_{NCVS}^{(i)}\), and the estimated sample variances, \(\hat{\sigma}_{NCVS}^{2}\).

\[
\hat{\alpha}_{(i)} = \hat{\mu}_{NCVS}^{(i)} \left( \frac{\hat{\mu}_{NCVS}^{(i)} (1 - \hat{\mu}_{NCVS}^{(i)})}{\hat{\sigma}_{NCVS}^{2} (i)} - 1 \right), \quad \hat{\beta}_{(i)} = (1 - \hat{\mu}_{NCVS}^{(i)}) \left( \frac{\hat{\mu}_{NCVS}^{(i)} (1 - \hat{\mu}_{NCVS}^{(i)})}{\hat{\sigma}_{NCVS}^{2} (i)} - 1 \right) \tag{5.6}
\]

Histograms for \(\hat{\mu}_{NCVS}^{(i)}\) and \(\hat{\sigma}_{NCVS}^{2}\) are presented in the top panel of Figure 15. We present the means and variances rather than the Beta parameters directly since it is difficult to interpret \(\hat{\alpha}_{(i)}\) and \(\hat{\beta}_{(i)}\). Both histograms are roughly bell-shaped, with some variation in parameter values. The county means are centered around roughly 0.67, meaning that the prior estimate of police reporting rate for most counties is close to 67%—which is equal to the estimated national police reporting rate for aggravated assault using the NCVS 2011 data. This is an indicator that the county-level prior distributions are consistent with published national trends, but remember that our numbers are not directly comparable to national estimates because of the non-random pattern of counties missing NIBRS data.
Figure 15: Histograms of prior and posterior county-level means and variances for the distribution of reporting rate

Plotting the beta distributions directly, however, shows that there is little variation in shape and location. Figure 16 displays a random sample of 100 county-level priors; the plot of all 1,446 priors looks very similar, but including all priors on the plot completely obscures individual lines. A concern is that the lack of variability among priors could cause the same smoothing effect that we found with the simple negative binomial model.
Figure 16: Sample of 100 beta distributions (out of 1,446 total) using the estimated parameters

Since the Beta distribution is the conjugate prior for the Negative Binomial, updating the prior is straightforward. If the prior distribution of police reporting rate for aggravated assault in county \( i \) is given by \( \text{Beta}(\alpha_{j(i)}, \beta_{j(i)}) \), then the corresponding posterior distribution is \( \text{Beta}(\alpha_{j(i)} + r, \beta_{j(i)} + x) \), where \( r \) is the number of observed “successes” and \( x \) is the number of observed “failures.” We can use the number of aggravated assaults reported to the police through NIBRS in county \( i \) as the number of observed successes, but we do not know the number of failures—that is, the number of aggravated assaults not reported to the police. The negative binomial model in Section 5.3, however, gives us an estimate of the total number of aggravated assaults in county \( i \) under the assumption that \( \rho_j = \rho_{j(i)} \) for each county \( i \) nested in a large area \( j \). Subtracting the number of NIBRS aggravated assaults from the estimate of total aggravated assaults from Section 5.3 will give us an estimate of the number of aggravated assaults not reported to the
police that is based on large-area geographic information from the NCVS. Again, we tested the model both using the means of the negative binomial distributions (rounded to the nearest whole number) and random draws from the distributions; results from each method agreed closely, so we show only the results based on using the means.

After updating, the histograms of the posterior means and variances showed much more variability than the prior parameters. This is largely due to a few counties with a high number of aggravated assaults reported through NIBRS. The posterior distributions show a greater variety of shapes and locations as well. The general shapes of most county-level distributions are still centered around 0.7 with the bulk of the density between 0.6 and 0.8, as was the case for the priors, but the shapes of the distributions are much more varied.

![Figure 17: Plot of beta posterior distributions of county-level police reporting rate (n=1,446)](image)
We include only the table for Ohio counties; the Midwest only and national tables showed similar trends, so they are not presented here for clarity. In general, the police reporting rates from this model are somewhat lower and show more variability than the police reporting rates from the negative binomial only model, but they are higher than the rates from the simulation model. Again, published NCVS estimates for the state of Ohio are not available, but the 2011 NCVS estimate for national police reporting rate was 67%, very close to the 68.41% predicted population-weighted rate for Ohio.

This procedure is an improvement over the negative binomial only model in Section 5.3, despite its simplicity. Adding the prior based on demographic county-level information allows for more variation in county-level aggravated assault police reporting rates, and does not force all counties in a geographic cell to have the same reporting rate. The posterior distributions also provide some information on means and on variability across counties—although there are no formal parameters, it is possible to use descriptive statistics to find broad trends in the county-level means and variances (e.g., which region has the highest mean reporting rate? Or, are reporting rates for urban areas more variable than in rural areas?). This procedure still does not incorporate the hierarchical structure of the data, however, and seems like it is not correctly indicating the county-level variability in police reporting rates.
The simulation-based method in Chapter 4 resulted in a 23 percentage point difference between the minimum and maximum county-level police reporting rates in Ohio; the equivalent police reporting rates from this procedure show only a nine percentage point difference. While we cannot guarantee that the simulation-based method is correct, the method was specifically designed to capture the variability of NCVS data to the extent possible; such a large difference from that method indicates that this procedure is probably underestimating the variability of police reporting rates, at least in part because the county-level priors are too similar.

In this application of the procedure, all beta distributions observed appeared to be unimodal. This is not necessarily the case: a beta distribution can have a wide variety of
shapes, including a U-shaped density, a bimodal density, or one with an asymptote at either zero or one. This poses problems for estimation, because without unimodality the posterior mean is not necessarily the area of highest density (or even with unimodality). The support of a beta distribution also does not contain zero or one, so this procedure would not be able to handle any cells with direct estimates of police reporting rate that are exactly zero or one. While this is not the case with the crime of aggravated assault for any of the cells used here, it is possible that a crime like rape (which is relatively rare and has a generally low police reporting rate) could have a direct estimate of zero for some cells. For all these reasons, we reject this procedure and instead turn to the procedure described in the next section.
Section 5.5: Hierarchical Bayes-based procedures and results

Method A: Single variance parameter

To remedy the issues above, we propose a procedure adapted from a Bayesian hierarchical model; call the procedure specified in Equation 5.7 “Method A.” Method A is very similar to the beta-negative binomial method in Section 5.4, but replaces the beta prior on estimated county-level police reporting rate with a normal prior with a logit link.

\[ Y_{j(i)} \sim \text{Negative Binomial} \left( Z_{j(i)} + 1, \rho_{j(i)} \right) \]
\[ \text{logit}(\rho_{j(i)}) = \mu_{j(i)} \]
\[ \mu_{j(i)} \sim \text{Normal} \left( \mu_{j(i)}^{\text{NCVS}}, \sigma^2 \right) \]

In Method A, as before, \( Z_{j(i)} \) is the number of observed aggravated assaults in county \( i \) nested in large area \( j \). Ideally, we would have direct county-level NCVS estimates to use for \( Y_{j(i)} \), the total number of all crimes in county \( i \), but this data is not available. Instead, we use county-level estimates of the number of crimes we would expect to observe under the strong assumption that all county-level police reporting rates \( \rho_{j(i)} \) are equal to the corresponding large-area NCVS police reporting rates \( \rho_j \). This allows us to incorporate information based on NCVS large area estimates into the model. Each \( Y_{j(i)} \) is set equal to the mean of \( \text{Negative Binomial} \left( Z_{j(i)} + 1, \rho_j \right) \). (See Section 5.4 for more details).
We take a logit transform of $\rho_{j(i)}$ so that we are making inference on the real line rather than on the interval $[0, 1]$. The middle step of the hierarchical model may seem unnecessary as written, but we separate the steps to emphasize that it is possible to add covariates in the second step. Under Method A the county-level police reporting rate $\rho_{j(i)}$ depends only on an overall mean parameter, $\mu_{j(i)}$, but it is possible to add other county-level covariates. For example, since the county-level police reporting rate might also depend on the percentage of residents in poverty and the county-level median income, then the second step in Equation 5.7 could be written as $\rho_{j(i)} = \mu_{j(i)} + \beta_1 * \text{poverty}_{j(i)} + \beta_2 * (\text{median income}_{j(i)})$. Adding covariates requires careful thought about an appropriate choice of prior for the overall mean term after the effect of all covariates is accounted for, so for simplicity we do not add covariates at this time.

We use the population-based county-level means derived in Section 5.4, $\mu_{j(i)}^{\text{NCVS}}$, as the mean for the normal prior distribution. Recall that these means are a population-based weighted average of NCVS police reporting rates for 12 sex, age, and race categories. The variance, $\sigma^2$, is considered common across all counties in the model. This parameter controls the level of importance we place on the prior estimate $\mu_{j(i)}^{\text{NCVS}}$. A small value of $\sigma^2$ indicates that the prior distribution is highly concentrated around the prior mean, and the geographically-based estimate will not have too much influence on the posterior distribution. A large $\sigma^2$ means the prior is relatively flat, and so the geographically based estimate $Z_{j(i)}$ will substantially influence the estimate of the posterior mean county-level police reporting rate. We place a uniform hyperprior on the precison of the prior, $1/\sigma^2$. 

163
WinBUGS, which is used for the model fitting, uses precision rather than variance in the model specification. We work with the precision here, although in future versions of this model we plan to use variance instead. The hyperprior placed on the precision is Uniform(0, 50). We performed sensitivity testing by varying the upper bound of the uniform distribution and found that the specific choice of upper bound had only a negligible effect on the final estimates, provided that the upper bound was greater than 25. (Values less than 25 seemed to truncate the posterior distribution of the precision.)

Method A and all subsequent models in this chapter were fitted using Markov chain Monte Carlo (MCMC) simulation in WinBUGS, a freeware designed to fit Bayesian hierarchical models, via the rube package in R. WinBUGS is widely used for fitting such models, such as in Law et al. (2014) and in Liu, Lahiri and Kalton (2007). WinBUGS model code for all models is provided in Appendix C. For each model, three parallel chains were run from the same starting values for 10,000 iterations each. The first 1000 iterations from each chain were discarded as burn-in and each chain was thinned by retaining only every tenth iteration, meaning that posterior estimates are based on a total of 900 kept iterations.

We assessed model convergence by visual inspection of the trace plots for each chain, autocorrelation plots, and the Gelman-Rubin plots. It was not possible to inspect the trace plots for the means of all 1,446 counties, so we reviewed plots for a sample of approximately 20-30 counties as well as any counties identified as outliers in summary
statistics. The `rube` output plots for the posterior distribution of the precision \( \tau \) (variable name tau.ncvs in the model code) are displayed in Figure 18 as an example; plots for selected county-level means can be found in Appendix D. The three color-coded lines in the figure represent the realizations of each chain. All three plots indicate the chains are mixing and converging acceptably well: the trace plot (top panel) does not display a trend, the autocorrelation plot in the middle panel shows that the thinned observations are not significantly autocorrelated after the first step, and the bottom panel indicates that the posterior distributions from all three chains are all roughly bell-shaped and agree reasonably well. The vertical lines in the center of the bottom plot denote the posterior means from each chain, which all occur around roughly 17.8; the other two sets of vertical lines identify the 95\% credible intervals for the posterior mean of each chain.

Figure 19 shows one summary diagnostic plot produced for the posterior estimates of the county-level means, \( \mu_{j(i)} \). The `rube` package randomly selects 20 parameters to display on the summary plot. The left panel displays the 95\% credible intervals for each chain (red, green, black) and the center plot displays the lag for which autocorrelation in the chain is significant at a 95\% confidence level. Short lags indicate a chain with draws that can be considered independent, and a more efficient simulation.
Figure 18: Diagnostic plots for checking convergence of precision parameter $\tau$ (tau.ncvs), Method A
The final plot on the right displays the Gelman-Rubin R-hat statistic for multiple chains; it is used to check convergence when multiple chains are run by comparing the within-chain and between-chain variances. Values of this statistic close to one typically indicate adequate convergence. The output plot is color-coded so that acceptable R-hat values are shown in green, borderline values are in orange, and high values are in red. Figure 19 tells us that while the 95% credible intervals for some posterior means are quite wide, autocorrelation does not seem to be a problem and the MCMC chains appear to converge properly. We checked summary plots for several hundred unique $\mu_{j(i)}$; all were similar to the figure below, so we do not include them here.

Figure 19: Summary diagnostic plot for checking convergence of selected county-level means (logit scale), Method A

167
Figure 20 compares the means and variances of the county-level posterior distributions of the police reporting rate $\rho_{j(i)}$ from Method A to the means and variances of the prior distributions and of the posterior distributions from the beta-binomial method in Section 5.4. Note that the “prior variance” for the beta-binomial model is the county-level variance derived using the NCVS replicate weights, and is only used in the beta-binomial method, not Method A. Also, because police reporting rate in Method A is on the logit scale, all posterior estimates generated by WinBUGS are also on the logit scale. The posterior means can easily be transformed back to the original scale simply by taking the inverse logit, but there is no simple way to reverse transform the posterior variances. Therefore, in Figure 20 the Method A means are comparable to the other means, but the Method A variances are not. Figure 20 also shows that Method A (labeled “HB”, or “hierarchical Bayes-based”) results in more variable posterior means than the beta prior method in Section 5.4. The increased variability is also evident in the histograms of posterior variance: the county-level posterior variances are much more spread out under Method A, to the point that the histogram is relatively flat, while posterior variances under the beta prior method are only slightly more spread out than the prior variances.
Figure 20: Histograms of county-level means and variances for prior distribution of $\rho_{(j)}$, posterior estimates for $\rho_{(j)}$ from beta-negative binomial model (Section 5.4), and posterior estimates for $\rho_{(j)}$ from hierarchical Bayes-based model (this section)
This increase in variability also seems to improve county-level estimates, as we anticipated. Table 26 and Table 27 compare the methods introduced thus far for Ohio counties. Method A has lower police reporting rates with a greater range of values than either the negative binomial only model or the beta-negative binomial method, which in turn results in somewhat higher and more variable estimates of county-level aggravated assault rate. The estimate of population-weighted mean aggravated assault rate for the state of Ohio under model A is still considerably lower than the estimate from the simulation-based method, however (1.60 per 1000 for model A versus 1.85 per 1000 under the simulation-based method). We take this as evidence that this method is promising, but likely requires improvement.

<table>
<thead>
<tr>
<th></th>
<th>Raw NIBRS</th>
<th>Simulation-based method (Chapter 4)</th>
<th>Negative binomial only model (Sec. 5.3)</th>
<th>Beta-negative binomial method (Sec. 5.4)</th>
<th>Hierarchical Bayes-based, method A (Section 5.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Min</strong></td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0180</td>
<td>0.0180</td>
<td>0.0272</td>
</tr>
<tr>
<td><strong>1st Qu.</strong></td>
<td>0.1647</td>
<td>0.3105</td>
<td>0.2530</td>
<td>0.2919</td>
<td>0.3112</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>0.3428</td>
<td>0.5907</td>
<td>0.4982</td>
<td>0.5466</td>
<td>0.5618</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>0.5185</td>
<td>0.8973</td>
<td>0.7438</td>
<td>0.7929</td>
<td>0.8254</td>
</tr>
<tr>
<td><strong>3rd Qu.</strong></td>
<td>0.5902</td>
<td>1.0740</td>
<td>0.8663</td>
<td>0.8958</td>
<td>0.9166</td>
</tr>
<tr>
<td><strong>Max</strong></td>
<td>2.4750</td>
<td>4.4150</td>
<td>3.4517</td>
<td>3.5067</td>
<td>3.7020</td>
</tr>
<tr>
<td><strong>Population weighted mean, Ohio</strong></td>
<td>1.0916</td>
<td>1.8455</td>
<td>1.4910</td>
<td>1.5253</td>
<td>1.5980</td>
</tr>
</tbody>
</table>

Table 26: Aggravated assault rates per 1000 people. Ohio only (n=86)
<table>
<thead>
<tr>
<th></th>
<th>Raw NIBRS</th>
<th>Simulation-based method (Chapter 4)</th>
<th>Negative binomial only model (Sec. 5.3)</th>
<th>Beta-negative binomial method (Sec. 5.4)</th>
<th>Hierarchical Bayes-based, method A (Section 5.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Min</strong></td>
<td>-</td>
<td>44.31%</td>
<td>58.66%</td>
<td>62.46%</td>
<td>65.03%</td>
</tr>
<tr>
<td><strong>1st Qu.</strong></td>
<td>-</td>
<td>53.41%</td>
<td>72.63%</td>
<td>65.88%</td>
<td>65.84%</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>-</td>
<td>57.84%</td>
<td>72.63%</td>
<td>66.34%</td>
<td>65.91%</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>-</td>
<td>57.30%</td>
<td>73.60%</td>
<td>66.68%</td>
<td>66.00%</td>
</tr>
<tr>
<td><strong>3rd Qu.</strong></td>
<td>-</td>
<td>60.59%</td>
<td>76.07%</td>
<td>67.06%</td>
<td>66.07%</td>
</tr>
<tr>
<td><strong>Max</strong></td>
<td>-</td>
<td>67.05%</td>
<td>78.90%</td>
<td>71.46%</td>
<td>67.75%</td>
</tr>
<tr>
<td><strong>Population weighted mean, Ohio</strong></td>
<td>-</td>
<td>59.19%</td>
<td>72.03%</td>
<td>68.55%</td>
<td>66.52%</td>
</tr>
</tbody>
</table>

Table 27: Aggravated assault police reporting rates in percentages. Ohio only (n=86)

*Method B: One variance parameter per MSA by region cell*

Recall that Method A specifies only one variance parameter across all counties. That is, the distribution of county-level aggravated assault police reporting rate is assumed to have a different mean for each county, but there is only one parameter that controls variance across all counties. A single variance parameter means that we are assuming that the weight placed on the large-area estimate versus the demographically-based estimate is exactly the same for all counties in the analysis. This is almost certainly not the case.

Letting variance vary by county is impractical, as it would require estimating another 1,446 parameters based on our already-sparse data. In Method B, we instead attempted to let the variance parameter vary by each MSA status by region cell. This method (Method B) can be written as
\[ Y_{j(i)} \sim \text{Negative Binomial} \left( Z_{j(i)} + 1, \rho_{j(i)} \right) \]

\[ \text{logit}(\rho_{j(i)}) = \mu_{j(i)} \]

\[ \mu_{j(i)} \sim \text{Normal} \left( \mu_{j(i)}^{\text{NCVS}}, \sigma^2 \right) \]

(5.8)

where \( j = 1, 2, \ldots, 12 \). Collapsing across place size ensures that each of the 12 cells has at least one NCVS observation. Method B was run in WinBUGS under an identical setup as Method A, using a Uniform(0, 50) hyperprior on all precision parameters and the same set of initial values. As under Method A, we tested different values for the upper bound of the uniform hyperprior on the precision and different initial values.

This model suffered from serious convergence problems, despite several adjustments to the bounds on the uniform hyperprior and the initial values. Method B resulted in errors on every run, and WinBUGS was never able to complete the full 900 kept iterations. The longest run resulted in 47 kept iterations; those are the results presented here. We suspect that some cells have too few NCVS victimizations for adequate estimation, since for each run of Method B we observed that the confidence intervals for certain cells expand as the upper boundary of the hyperprior increases. For example, the rows labeled [3,3] and [3,4] in Figure 21, which correspond to the rural South and rural West cells respectively, seem to be truncated by the upper boundary of the hyperprior. We find that this occurs for these cells even when the upper boundary is set to an extremely large value like 500 or 1000. The estimated precision for rows [2,1], [2,2], and [3,2] in the same figure, corresponding to the suburban Midwest, suburban Northeast, and rural Northeast, all have unacceptably wide 95% credible intervals for the posterior mean that also grow...
when the upper bound on the hyperprior is increased. These wide intervals are probably
due to the low number of counties in these cells (see Table 28).

Figure 21: Diagnostic plots for checking convergence of $\tau$ (tau.ncvs), Method B

<table>
<thead>
<tr>
<th></th>
<th>Midwest</th>
<th>Northeast</th>
<th>South</th>
<th>West</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban</td>
<td>203</td>
<td>42</td>
<td>226</td>
<td>94</td>
<td>565</td>
</tr>
<tr>
<td>Suburban</td>
<td>76</td>
<td>6</td>
<td>138</td>
<td>27</td>
<td>247</td>
</tr>
<tr>
<td>Rural</td>
<td>276</td>
<td>8</td>
<td>245</td>
<td>105</td>
<td>634</td>
</tr>
<tr>
<td>Total</td>
<td>555</td>
<td>56</td>
<td>609</td>
<td>226</td>
<td>1446</td>
</tr>
</tbody>
</table>

Table 28: Count of number of counties per cell, with index and MSA/region labels
Since $\tau_j$ is a precision parameter equal to $1/\sigma_j^2$, large estimated values of $\tau_j$ correspond to small estimated $\sigma_j^2$. This means that Model B is overstating the importance placed on the prior estimate $\hat{\mu}_{j(i)}^{NCVS}$ and understating the importance placed on the large-area based estimate $\hat{\mu}_{j(i)}$. The plot in Figure 22 is for the precision in the Northeast for suburban areas (one of the cells with few observations) to demonstrate the poor convergence properties of this model. The jaggedness of the trace plot is because this particular simulation stopped after 47 thinned iterations with a 1000 iteration burn-in (1470 total iterations). It is also evident that the three chains (red, green, and black) do not agree well on the posterior mean. Because of these serious issues, we reject Method B.

Figure 22: Diagnostic plots for precision parameter for suburban Northeast, Method B
**Method C: One variance parameter per MSA status**

We attempt to remedy some of these issues in Method C by collapsing across region, so that the standard deviation $\sigma^2$ depends on MSA status only:

$$\hat{Y}_{j(i)} \sim \text{Negative Binomial} \left( Z_{j(i)} + 1, \rho_{j(i)} \right)$$

$$\text{logit}(\rho_{j(i)}) = \mu_{j(i)}$$

$$\mu_{j(i)} \sim \text{Normal} \left( \mu_{NCVS}^{j(i)}, \sigma_{m}^2 \right)$$

where $m=1, 2, 3$ indexes the MSA status of cell $j$ (Central City MSA, non-Central City MSA, non-MSA). Again, Method C was run in WinBUGS under an identical setup as Method A, using a Uniform(0, 50) hyperprior on all variance parameters and the same set of initial values, with a variety of values tested for the upper bound of the hyperprior and different initial values. Method C is an improvement over Methods A and B in that the MCMC chains converged properly, yet it still allows some differences in county-level variance parameters based on MSA status. However, we have some concerns about the performance of this model as well.

In Figure 23, the ordering for the MSA categories is [1]=Urban, [2]=Suburban, [3]=Rural. The 95% credible interval for the posterior mean of the precision parameter $\tau_m$ for urban areas (1) is unremarkable, but the 95% credible interval for suburban areas (2) is very narrow and centered around a very small mean—only 2.134. A precision of 2.134 translates to a standard deviation of 0.685, an unacceptably large number when the median estimate for $\mu_{j(i)}$ is around 0.8. The 95% credible interval for the posterior mean of the precision parameter for rural counties has the opposite problem. As was the case
under Method B, we find that the credible interval tends to expand and the estimate for the posterior mean moves towards the upper limit of the hyperprior.

Figure 23: Diagnostic plots for checking convergence of precision parameter (tau.ncvs), Method C

We suspect that this is due to the large number of rural counties with zero police-reported aggravated assaults. 82 out of 634 (12.9%) total counties identified as rural have zero NIBRS aggravated assaults, but only 17 out of 247 (6.9%) suburban counties and 8 out of 565 (1.4%) of urban counties. The estimated standard deviation for rural counties is relatively small, since most rural counties have very few aggravated assaults as shown in
Table 29, but the estimate is also very unstable because with so few yearly assaults per county a difference of one or two aggravated assaults per county could seriously impact the final estimate.

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1st Qu</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu</th>
<th>Max</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban</td>
<td>0.00</td>
<td>27.00</td>
<td>71.00</td>
<td>285.90</td>
<td>225.00</td>
<td>11680.00</td>
<td>749.81</td>
</tr>
<tr>
<td>Suburban</td>
<td>0.00</td>
<td>4.00</td>
<td>16.00</td>
<td>50.62</td>
<td>46.00</td>
<td>841.00</td>
<td>111.01</td>
</tr>
<tr>
<td>Rural</td>
<td>0.00</td>
<td>2.00</td>
<td>7.00</td>
<td>16.77</td>
<td>18.00</td>
<td>321.00</td>
<td>30.58</td>
</tr>
</tbody>
</table>

Table 29: Summary statistics for number of NIBRS aggravated assaults per county by MSA status (n=1446)

Methods D and E: One variance parameter per region

We tried allowing the county-level variance to depend on region only instead (Method D); this model is nearly identical to Method C, except use $\tau_r$ where $r=1, 2, 3, 4$ and indexes region, rather than $\tau_m$. We again noticed some issues with the estimates for the precision parameter. In Figure 24, the first two regions (Midwest [1] and Northeast [2]) have very small estimates for $\tau_r$ and therefore large estimated standard deviations, while the last two regions (South [3] and West [4]) have relatively large posterior estimates for precision and therefore small estimated standard deviations. The Northeast has few counties in our data as compared to the other regions, so it makes sense that the Northeast would have a relatively large standard deviation as well (see Table 28 for exact counts of counties by region). But this is not the case for the Midwest; with 555 counties, it has almost ten times more counties in our dataset than the Northeast.
In the NIBRS 2011 data, there is one county that is an extreme outlier in terms of number of aggravated assaults, and this county happens to be in the Midwest: Wayne County, MI, which contains Detroit. Wayne County contains 11,681 aggravated assaults in NIBRS 2011; the next closest county is Shelby County, TN (in the South region, containing Memphis) with 6,784 aggravated assaults, and after that the count drops off to around 3,000 per county. Lincoln County, WV is also an extreme outlier in terms of police-reported aggravated assaults rate, with 49 per 1000 (the next largest is a rate of 16.92 per 1000); although West Virginia is in the South census region, we thought that removing such an extreme outlier might improve the overall performance of the model. We ran Method D with these two counties (Wayne County, MI and Lincoln County, WV) excluded; call this Method E. Removing the outliers had very little impact on the final
parameter estimates. For example, the four posterior means of the precision parameter for all counties in Method D were 5.03, 1.69, 35.88, & 34.76; the corresponding numbers from Method E were 4.95, 1.68, 36.06, & 33.47. Because the results from Method E were so similar to those from Method D, we include them in Appendix D.

The issue for the Midwest is more likely that it has so many counties with a NIBRS aggravated assault rate of 0, just as for rural counties in Method C. The histograms below show the number of counties reporting 10 or fewer aggravated assaults in NIBRS 2011 by region. The cluster of counties at zero aggravated assaults for the Midwest cause the standard deviation for number of county-level aggravated assaults to be larger, and therefore result in a smaller value of $\tau_r$.

Figure 25: Counties with fewer than 10 NIBRS aggravated assaults, by region
We also found that under Method D, $\tau_r$ for the West region tended to creep towards the upper boundary. However, Figure 24 also indicates that the posterior means for the precision parameters for the Northeast and Midwest are similar, as are the posterior means for the South and the West. Perhaps grouping regions together will lead to a more stable model.

*Method F: Two variance parameters assigned by region*

In Method F, we assigned one precision parameter $\tau_1$ to all counties from the Midwest and Northeast, and a second precision parameter $\tau_2$ to all counties from the South and West. This has a stabilizing effect on both precision parameters as seen in Figure 26, and we consider this the most successful model. More output plots for all models, including trace plots for selected parameters, are available in Appendix D.

Figure 26: Diagnostic plots for checking convergence of precision parameter (tau.ncvs), Method F
Table 30 compares the results from all six methods in this section. The first two rows present the average posterior mean and average posterior variance for all counties used in the method; a total of 1,446 counties for Method D and 1,448 counties for all other models. (Recall that the posterior means and variances are on the logit scale.) The next two rows provide the mean estimated police-reporting rate, calculated by taking a population-weighted average of the posterior means for each county in the model, and the mean estimated aggravated assault rate, calculated by taking a population-weighted average of the estimated county-level aggravated assault rates. The estimated population-weighted county-level aggravated assault rates are calculated based on the population-weighted police reporting rates exactly as in Section 5.4.

As stated in Chapter 4, averages of aggravated assault rates over all counties using this data are unreliable at best and misleading at worst because of the high percentage of counties missing NIBRS data in a pattern that cannot be considered MCAR. These “all counties” numbers are useful as comparisons between methods, but we urge great caution in interpreting them as inference about the national police reporting rate or national aggravated assault rate. The next two rows are the same two quantities for the Midwest, which we consider somewhat more reliable, and the following two rows are for the state of Ohio only. The final two rows of Table 30 contain the model estimates for Franklin County, Ohio alone, to demonstrate the changes in a single county’s estimates across the six methods. (Franklin County is an urban county in the Midwest, and the home of The Ohio State University.)
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average posterior mean (logit scale)</td>
<td>0.7401</td>
<td>0.8140</td>
<td>0.8093</td>
<td>0.7945</td>
<td>0.8242</td>
<td><strong>0.8204</strong></td>
</tr>
<tr>
<td>Average posterior variance (logit scale)</td>
<td>0.1850</td>
<td>0.2358</td>
<td>0.2381</td>
<td>0.1995</td>
<td>0.2003</td>
<td><strong>0.2215</strong></td>
</tr>
<tr>
<td>Mean est police reporting rate, all counties</td>
<td>67.62</td>
<td>69.06</td>
<td>69.03</td>
<td>68.69</td>
<td>69.26</td>
<td><strong>69.22</strong></td>
</tr>
<tr>
<td>Mean est agg assault rate, all counties</td>
<td>2.29</td>
<td>2.28</td>
<td>2.26</td>
<td>2.29</td>
<td>1.83</td>
<td><strong>2.25</strong></td>
</tr>
<tr>
<td>Midwest counties only, pop-weighted est police reporting rate</td>
<td>69.97</td>
<td>70.00</td>
<td>70.38</td>
<td>69.39</td>
<td>69.62</td>
<td><strong>71.05</strong></td>
</tr>
<tr>
<td>Midwest counties only, pop-weighted est agg assault rate</td>
<td>3.02</td>
<td>3.03</td>
<td>3.01</td>
<td>3.10</td>
<td>3.09</td>
<td><strong>3.00</strong></td>
</tr>
<tr>
<td>Ohio counties only, pop-weighted est police reporting rate</td>
<td>69.32</td>
<td>69.40</td>
<td>70.38</td>
<td>69.85</td>
<td>68.81</td>
<td><strong>70.92</strong></td>
</tr>
<tr>
<td>Ohio counties only, pop-weighted est agg assault rate</td>
<td>1.52</td>
<td>1.52</td>
<td>1.51</td>
<td>1.52</td>
<td>1.53</td>
<td><strong>1.51</strong></td>
</tr>
<tr>
<td>Franklin County estimated police reporting rate</td>
<td>71.61</td>
<td>75.31</td>
<td>71.58</td>
<td>70.04</td>
<td>70.07</td>
<td><strong>72.31</strong></td>
</tr>
<tr>
<td>Franklin County estimated agg assault rate</td>
<td>1.84</td>
<td>1.75</td>
<td>1.84</td>
<td>1.88</td>
<td>1.88</td>
<td><strong>1.82</strong></td>
</tr>
</tbody>
</table>

Table 30: Comparison of all six methods from Section 5.5

The variation between methods is relatively minor, but in a pattern we expect. The average posterior variance is lowest in Method A, in which all county-level variances are assumed equal, and highest in Method B, where we allowed twelve different parameters for county-level variance. Our chosen method, F, has an average posterior variance right
in the middle of all methods considered. Method D appears to have a much lower estimate for aggravated assault rate among all counties, but most of that drop is due to excluding Detroit—this number is actually comparable to the other models if Wayne County is excluded (the county in West Virginia excluded has a very small population, and therefore very little impact on the population-weighted aggravated assault rate.) We consider it important to retain Wayne County in the final model: large urban areas tend to be where the most crime occurs, and excluding a county because of high crime counts will cause us to underestimate the overall crime rate. Notice also that the estimated police-reporting rate for Method D is comparable to the other five methods.

Figure 27 compares estimates for Ohio counties across three methods: the large area only model used in Section 5.3, the beta prior method used in Section 5.4, and Method F discussed above. The estimates from all three methods appear to be quite similar; in fact, the maps for the beta prior method and Method F are identical. This is due in part to the coarseness of the categories on the map, which mask small changes in crime rates. But it may also be an indication that the limited data available will result in similar county-level estimates regardless of the modeling strategy used. This fact paired with the lack of reliable gold-standard estimates for comparison makes it extremely difficult to select a “best” method based on final county-level estimates alone.
We find that of all the methods tested in this chapter, a method adapted from a hierarchical Bayes model as specified at the beginning of this section with two variance parameters defined by region (Model F) is the most successful. It fulfils nearly all of the criteria specified in Section 5.1: it is relatively simple, takes advantage of the hierarchical structure of the data, includes variance parameters, and is widely generalizable. The
procedure may be overly simplistic, but without additional county-level information from the NCVS we find it difficult to improve this method further.
Chapter 6: Conclusions and directions for future research

An often-repeated quip among crime researchers is that there is just never enough crime. From nearly every other perspective low crime rates are desirable, but when working with crime data its sparsity poses serious challenges. The problem we address in this research—modeling crime rates at the county level in a way that accounts for all crimes, reported and unreported—is especially challenging because of the lack of reliable data at the county level. We cannot rely on the NCVS alone because although it is the most reliable source for data on all crimes, county-level identifiers are not publicly available for victimization records in the NCVS due to privacy concerns. Our research attempts to link data sources at two different levels (NIBRS data is available at the county level but NCVS victimization data only provides broad geographic identifiers like region, MSA status, and place size) with the added complication that the two sources record overlapping but not identical subsets of crime. NIBRS is designed to count police-reported crimes only from all victims, regardless of age and including non-individual victims like businesses; the NCVS records both reported and non-reported victimizations, but excludes children under 12 and non-individual victims.

In Chapters 4 and 5, we have presented two relatively simple and widely accessible methods for combining NIBRS data from police reports with the publicly available
Neither method is computationally intensive: for the simulation-based method in Chapter 4, we were able to simulate 50,000 draws for each of over 3,000 counties in about 6 hours using R’s sample package, and all of the methods proposed in Chapter 5 ran in less than five minutes in WinBUGS, both on a Quad Eight Core Xeon 2.13 with 128GB RAM. Results from both methods applied to 2011 NIBRS and NCVS aggravated assault data are easily interpretable and agree with expected aggravated assault trends; for example, we consistently find higher crime rates in urban counties. By using one year of data only, we maintain the ability to detect year-to-year variation in county-level crime trends if these methods are used to predict county-level crime rates for multiple years. (Pooling four years of NCVS data to estimate police-reporting rates in the simulation-based method in Chapter 4 will have a smoothing effect on trends in police-reporting rates, but using only one year of NIBRS data should preserve most year-to-year variation.)

A criticism of both methods is that they are too simplistic and may not adequately account for the complex social and economic factors that drive crime rates. We believe that the available data, especially from the NCVS, is too sparse to support more complex modeling, and that we risk overfitting to the 2011 data if we incorporate too many covariates. For example, when fitting the hierarchical Bayes-based procedures in Section 5.5, Method B failed to converge properly with only twelve variance parameters. We are also limited in our choice of covariates by missing data in the NCVS. Previous research on crime trends indicates that poverty and transience are important variables to
consider—poor areas are likely to have higher crime rates, as are areas where residents move frequently. The NCVS data is missing income and length at current residence data for approximately half of victimization records in the 2011 file—a percentage so large that we cannot use most standard imputation techniques or simply drop cases with missing information, especially since income and transience information are typically not missing at random. Persons who move frequently are more likely to forget the date of their last move, or refuse to give the information at all; they are also more likely to be missed by NCVS interviewers altogether, since they may move in and out of a housing unit between field interview periods. Poverty tends to be correlated with transience, and the very poor and the very rich typically refuse to provide income information at a higher rate than middle income households. For these reasons, we cannot use poverty status or transience in our models despite the strong and established influence of these factors on crime rates.

Our procedures are also designed to be as widely applicable as possible to different crime types and different small areas; one could consider states as small areas or even small demographic subgroups, and change the covariates accordingly. We intend for the methods presented in this research to be intuitive enough that crime researchers would be able to apply them to their own areas of interest, even those who are not necessarily trained statisticians.
Future research includes refining both methods with access to the NCVS county-level identifiers and information on NCVS strata groupings. If we could assign NCVS cases to particular counties and knew which counties BJS considered similar for sampling purposes, we could incorporate that information into our sampling procedure or as a level in the hierarchical model rather than the rough region/MSA categories we are currently using. We chose not to pursue access to the county-level identifiers for this research, however; since the identifiers are not publicly available, our methods would not be easily replicable by other researchers.

A weakness of our methods is the limited NIBRS data coverage. We cannot make predictions for counties in which no law enforcement agencies report to NIBRS, which as of this writing is more than half of all U.S. counties. BJS has made it clear, however, that they view NIBRS as the future of police-reported crime statistics, so it makes sense to rely on NIBRS to develop methods with the assumption that NIBRS coverage will continue to improve over the next several years. Another avenue for future research is an extension of both methods to rely on UCR data rather than NIBRS data. We use NIBRS information to easily aggregate crimes to the county level and to exclude crimes committed against non-individual victims or victims under age 12. It is relatively straightforward, albeit more tedious, to aggregate UCR data to the county level. Addington (2007) suggests a method for estimating the percentage of UCR crimes committed against victims outside the scope of the NCVS. We chose not to use counties with UCR only data in this research because we did not want to add another layer of
estimation to our modeling, but it is possible to extend this research to all counties with UCR data. Estimates from such a model should be used with extreme caution, however—they would be much more reliable estimates of crime trends (e.g., is crime going up or down?) than of absolute crime rates.

As mentioned in earlier chapters, we also cannot easily assess how accurate the estimates produced by each method are, since there are currently no county- or state-level gold standards for comparison, and our national estimates are unreliable due to the number of missing counties. Fortunately, BJS is performing a pilot study to improve NCVS state-level estimation by increasing sample in 11 states (including Ohio) to permit direct estimation, and plans to have 22 states with enough NCVS sample cases for direct estimation by 2016 (Planty 2014). Our current estimates are almost certainly an underestimation because we do not account for NIBRS crimes reported to agencies that cover multiple counties such as state police forces. NCVS state-level estimates would help us assess the impact of excluding such crimes, and let us test methods for allocating such crimes to counties.

Although these methods were developed using crime data and designed for county-level estimation, they are generalizable to a wide variety of other statistical problems. Survey data is expensive to collect, especially when observing rare events like crime, and high-quality survey data often has limited coverage or even no direct sample at all in small areas of interest. Administrative data is widely available in many applications, but since it
is typically collected for some other purpose it may not directly cover the population of interest. For example, in the health care field, administrative data on patient records is available through the Centers for Medicare and Medicaid Services (CMS), but it is limited to only persons enrolled in Medicare and Medicaid. Medicare and Medicaid patients are different from the general population of the United States: these programs are designed to cover persons over 65, the low-income, persons with certain medical conditions or disabilities, and pregnant women and children who would otherwise be uninsured. The National Health Interview Survey (NHIS) collects detailed health information about all respondents, but the geographic information available is limited to respect respondents’ confidentiality. A researcher interested in prevalence of a relatively rare disease like Hepatitis C at the county level could adapt our methods to combine available Medicare/Medicaid data for the counties of interest with detailed NHIS survey records. The growing availability of large administrative datasets paired with interest in estimates for small areas suggests that there are similar problems in many other applications as well that could benefit from the methods proposed here.

Finally, in Chapter 5 we propose a novel adaptation of Bayesian modeling strategies by using county-level estimates based on NCVS data rather than observed county-level data to update the county-level priors. This type of procedure could also be modified for other statistical problems for which Bayesian strategies would be effective, but response data for each unit of analysis is not available. It can be considered an extension of empirical Bayes methods, in which both parameters and data can be replaced with estimates based
on observed data. Such a method could also be useful for applications where the variable of interest is continuous but measured on a coarse categorical scale: for example, suppose that a researcher is interested in estimating household income, but the available survey data only asks respondents whether they fall in a “low,” “medium,” or “high” income group. She could construct a prior distribution for each respondent’s income based on demographics, then update that prior with some other estimate of income based on the survey question and some other respondent characteristics.

The problem of estimating crime rates accounting for unreported crime at the small area level is difficult and complex, and we do not claim that our methods are a complete solution. They are, however, a clear improvement over previous research using only police-reported crime data, and over methods that use only the aggregate UCR counts with the NCVS data. They take advantage of the particular strengths of each data source—the rich county-level data available through NIBRS and information on crime not reported to the police in the NCVS—while minimizing the impact of each source’s weaknesses. Though developed specifically for estimating county-level crime rates, our methods are also applicable to other statistical problems and data sources.

Both the NCVS and NIBRS are in the process of increasing data quality through sample increases and improvements in data collection processes, which will improve estimates made via the methods proposed here or related methods. Public interest in reliable crime statistics, especially for crimes related to vulnerable populations like women, children,
and inmates, has led to increased federal and state investment in BJS and the NCVS, as well as in NIBRS and the UCR program. Perhaps improved data paired with methods such as ours that borrow strength across data sources will soon mean that reliable inference can be made even if there is “not enough crime.”
References


Appendix A: Derivation of appropriate negative binomial distribution

We are interested in the distribution of the total number of trials \( n \) required to observe \( s \) successes, given a certain probability of success \( p \). (Or in context, \( n \) is the total number of crimes, which is unknown; \( s \) is the number of NIBRS crimes, which is known; and \( p \) is the police reporting rate, which we estimate.) The negative binomial distribution seems like the appropriate choice, but it provides the distribution of \( n \) only in the special case when the final trial observed is a success. That is, if it is possible to somehow observe trials in order and mark down the total number of trials as \( n \) when you observe the \( s \)th success. That is not the case here, since we have no ordering imposed on the crimes and thus no guarantee that the “last” crime will be a police-reported crime. Intuitively, what we are interested in is actually the total number of trials before observing the \( s+1 \)th success—we can observe \( s \) successes and then any number of failures, but we must stop before the next success—which would be a Negative Binomial\( (s+1, p) \) distribution.

It is possible to prove this quantitatively using a Bayesian argument. The following discussion draws heavily on Chapter 8 of Vose (2008). Let \( x \) be the number of failures that occurred before the \( s \)th success for some fixed value of \( p \). Then place a uniform prior on \( x \), since we have no information about \( x \) but we can make an educated guess about its range: \( p(x|p) = c, \ x \in (0, \frac{1}{c}) \).
We do not know the total number of trials, but we do know that $s$ trials were successful with probability of success $p$. Therefore, the likelihood function for the number of failures $x$ given $s$ and $p$ is just proportional to a binomial probability mass function:

$$p(x|s, p) \propto \binom{s + x}{s} p^s (1 - p)^x$$

Bayes’ Theorem tells us that:

$$p(x|s, p) = \frac{p(s|x, p)p(x|p)}{p(s|p)}$$

We know that $p(s|x, p)$ is simply a binomial pmf with $n=x+s$, and we have assumed that $p(x|p)=c$ for some fixed value of $c$. We need to find the normalizing constant, $p(s|p)$.

$$p(s|p) = c \sum_{i=1}^{\infty} \binom{s + i}{s} p^s (1 - p)^i$$

This sum turns out to be simply $c/p$. We can then simplify:

$$p(x|s, p) = \frac{\binom{s + x}{s} p^x (1 - p)^x \cdot c}{c/p} = \binom{s + x}{s} p^{s+1} (1 - p)^x$$

which is simply the pmf of Negative Binomial($s+1$, $p$).

201
Finally, if $X_r$ is a random variable distributed as Negative Binomial($r, p$), then it can be shown that $X_r$ is a sum of $r$ independent Geometric($p$) variables. $X_r$ is therefore approximately normal for sufficiently large $r$ by the Central Limit Theorem; we use this fact when constructing confidence intervals involving negative binomial variables.
Appendix B: Map of Ohio counties, labeled with county names
Appendix C: WinBUGS model code

This appendix includes the WinBUGS model code only. All R code, including the code used to set initial values and to execute the model, is available from the author upon request.

Note: R and WinBUGS parameterize the negative binomial as the distribution of the number of failures observed before s successes, so all the coding is done under that parameterization—meaning that the mean number of failures is given by \((Y_j(i) + 1) \times \left(\frac{p_j}{1-p_j}\right)\), which is exactly the same as the formula above minus \(Y_j(i) + 1\).

WinBUGS also uses precision, not variance, to describe distributions. For example, in model 1 the line \(\text{dnorm(mu.ncvs[i], tau.ncvs)}\) specifies a normal distribution with mean mu.ncvs[i] and variance 1/tau.ncvs. More information on WinBUGS can be found in the WinBUGS User Manual, Spiegelhalter et al. (2003).

Method A

\[
\begin{align*}
\text{ml}<"\text{model}\{
\text{for } (i \text{ in } 1:N)\{\\n\text{Z[i]~dnegbin(p[i], Y[i])}\\n\text{logit(p[i])<-mu[i]}\\n\text{mu[i]~dnorm(mu.ncvs[i], tau.ncvs)}
\}
\}
\end{align*}
\]
Method B

m2<-"model{

  for (r in 1:R){
    for (m in 1:M){
      for (i in 1:N[m, r]){  
        Z[i, m, r]~dnegbin(p[i, m, r], Y[i, m, r])
        logit(p[i, m, r])<-mu[i, m, r]

        mu[i, m, r]~dnorm(mu.ncvs[i, m, r], tau.ncvs[m, r])
      }

      tau.ncvs[m, r]~dunif(0, 40)
    }
  }
}

""

Method C

##by MSA
m3<-"model{

  for (m in 1:M){
    for (i in 1:N[m]){  
      Z[i, m]~dnegbin(p[i, m], Y[i, m])
      logit(p[i, m])<-mu[i, m]

      mu[i, m]~dnorm(mu.ncvs[i, m], tau.ncvs[m])
    }

    tau.ncvs[m]~dunif(0, 50)
  }

}"
Models D, E, and F

##by region for model D
##same model was run for Method E, with the noted observations dropped from the dataset
##also modified for grouped region, just let R=2 rather than R=4.

```r
m4<="model{
  for (r in 1:R){
    for (i in 1:N[r]){[code]
      Z[i, r]~dnegbin(p[i, r], Y[i, r])
      logit(p[i, r])<-mu[i, r]

      mu[i, r]~dnorm(mu.ncvs[i, r], tau.ncvs[r])
    }
  }
  tau.ncvs[r]~dunif(0, 50)
}
"
```
Appendix D: Selected diagnostic plots from Chapter 5 model output

Method A

Figure 28: Diagnostic plot for deviance, Method A
Figure 29: Diagnostic plot for posterior for Franklin County, OH under Method A
Figure 30: Summary diagnostic plot for selected county-level means, Method B

Figure 31: Diagnostic plot for deviance, Method B
Figure 32: Diagnostic plot for tau.ncvs (variance parameter by large area), Method B

Figure 33: Diagnostic plot for selected tau.ncvs (variance parameter for rural Midwest), Method B
Figure 34: Diagnostic plots for posterior for Franklin County, OH under Method B
Method C

Figure 35: Summary diagnostic plot for selected county-level means, Method C

Figure 36: Diagnostic plot for deviance, Method C
Figure 37: Diagnostic plot for tau.ncvs by MSA status, Method C

Figure 38: Diagnostic plot for tau.ncvs[1] (variance parameter for urban counties), Method C
Figure 39: Diagnostic plot for tau.ncvs[2] (variance parameter for suburban counties), Method C

Figure 40: Diagnostic plot for tau.ncvs[3] (variance parameter for rural counties), Method C
Figure 41: Diagnostic plots for posterior for Franklin County, OH under Method C
Method D

Figure 42: Summary diagnostic plot for selected county-level means, Method D

Figure 43: Diagnostic plot for deviance, Method D
Figure 44: Diagnostic plot for tau.ncvs (variance parameter by region), Method D

Figure 45: Diagnostic plot for tau.ncvs[1] (variance parameter for Northeast), Method D
Figure 46: Diagnostic plot for tau.ncvs[2] (variance parameter for Midwest), Method D

Figure 47: Diagnostic plot for tau.ncvs[3] (variance parameter for South), Method D
Figure 48: Diagnostic plot for tau.ncvs[4] (variance parameter for West), Method D
Figure 49: Diagnostic plots for posterior for Franklin County, OH under Method D
Method E

Figure 50: Summary diagnostic plot for selected county-level means, Method E

Figure 51: Diagnostic plot for deviance, Method E
Figure 52: Summary diagnostic plot for tau.ncvs, Method E

Figure 53: Diagnostic plot for tau.ncvs[1] (variance parameter for Northeast), Method E
Figure 54: Diagnostic plot for tau.ncvs[2] (variance parameter for Midwest), Method E

Figure 55: Diagnostic plot for tau.ncvs[3] (variance parameter for South), Method E
Figure 56: Diagnostic plot for tau.ncvs[4] (variance parameter for West), Method E
Figure 57: Diagnostic plots for posterior for Franklin County, OH under Method E
Method F

Figure 58: Summary diagnostic plot for selected county-level means, Method F

Figure 59: Diagnostic plot for deviance, Method F
Figure 60: Diagnostic plot for tau.ncvs[1] (variance parameter for Northeast and Midwest), Method F

Figure 61: Diagnostic plot for tau.ncvs[2] (variance parameter for South and West), Method F
Figure 62: Diagnostic plots for posterior for Franklin County, OH under Method F