Small Area Estimates from the National Crime Victimization Survey

Robert E. Fay[1], Michael Planty[2], Mamadou S. Diallo[1]
[1]Westat, Inc., 1600 Research Boulevard, Rockville, MD 20850
[2] Bureau of Justice Statistics; 810 Seventh Street, NW; Washington, DC 20531

Abstract
The National Crime Victimization Survey (NCVS) has provided annual estimates of criminal victimizations for four decades, almost exclusively focusing on national results. Previous papers have described an effort to develop small area approaches to produce estimates for states, and for large counties and cities. In this paper, we report the results from this effort and compare the estimates to the ongoing Uniform Crime Report (UCR) of the FBI. We describe methodological refinements to the methods for time-series modeling proposed by Rao and Yu. We also summarize substantive implications of the new results for understanding the geographic distribution of crime in the United States.

Key Words: NCVS, Rao-Yu model, violent crime, multivariate Rao-Yu model.

1. Introduction

The National Crime Victimization Survey (NCVS) is an ongoing national survey of the civilian, non-institutional population in the U.S. The survey produces annual estimates of crime as reported by its victims. The survey universe is the civilian, non-institutional population age 12 and over. The data for victimization by violent crime—rape and sexual assault, aggravated assault, simple assault, and robbery—are provided primarily by self-response. Data on property crime—burglary, motor vehicle theft, and other theft—are provided by household respondents. The Bureau of Justice Statistics (BJS) sponsors the survey, which is conducted by the Census Bureau.

The NCVS sample design is a multi-stage sample of housing units and non-institutional group quarters. Basic aspects of the sample design affect the use of the data for small area estimation. In particular, the sample design results in sampling correlations among the estimates over time for two reasons. First, the survey employs a rotating panel design, with sampled households remaining in sample for a total of 7 interviews scheduled 6 months apart. Respondents report on crimes during the previous 6 months. Households who move from a sampled housing unit are not followed, but instead they are replaced by incoming households. Second, the first-stage design of the survey historically has been redesigned on a ten-year cycle, so most adjacent years are estimated based on the same first-stage sample of primary sampling units. An additional complication has been recently introduced: Until the last few years, the first interview was excluded from the published estimates, when its purpose was to bound the second interview with a more definitive timeframe. Research has demonstrated that respondents in unbounded interviews tend to telescope events into the reference period leading to over-reporting. Since 2006, as a cost saving measure due to reductions in sample sizes, the data from the

unbounded first interview has instead been statistically adjusted and included in the estimation.

In the past, the geographic detail available from the survey has been quite limited. BJS publishes an annual bulletin and special reports with crime statistics analyzed by major demographic domains, but with typically little geographic detail. NCVS microdata files are publicly available, but they contain geographic information limited to the four regions; for example, state of residence is unavailable. The confidentiality of the NCVS data is protected under Title 13, the enabling legislation for the Census Bureau, and restricting geographic detail has been one of the primary strategies to mitigate disclosure risk. Further, the sample design is not stratified by state or smaller geographic area such as city or county. The result is a sample that may not represent a specific area (e.g., state) or lacks the necessary sample size to produce reliable estimates.

The occurrence of crime, the enforcement of the criminal statutes, and determination of policing policy are all primarily or exclusively local rather than national issues. For that reason, BJS is supporting research to extend the utility of the NCVS by expanding its geographic detail. BJS has released a limited set of subnational victimization estimates in various reports utilizing existing data when available (e.g., Lauritsen and Schaum, 2005) or from entirely new collections (Smith et al. 1999). The value of subnational estimates of criminal victimization is three-fold: description, evaluation, and allocation. Subnational estimates can be used to describe local crime problems and trends that can vary substantially from national patterns. Victimization estimates are independent from the reporting and processing filters associated with law enforcement measures and can highlight the "dark figure" of unreported crime (Skogan 1977). Unlike the variation is local laws and police practices, the NCVS provides a vehicle to compare standardized measures of crime across geographic areas and to the nation as a whole. Further, relying solely on police statistics may underestimate the true level and nature of crime, as many personal crimes such as sexual assault and domestic violence are not reported to police (Truman and Planty, 2012). Other crimes go unreported because the victim feels the police would not or could not help, deal with it another way, or are afraid of reprisal (Langton et al. 2012). Victimization statistics can provide local authorities a measure of understanding as to why or why not citizens contact and cooperate with the police.

Secondly, having more information about the spatial variations of the crime problem in relation to other social, economic, and political factors are essential to understanding the mechanisms related to the level and change in crime. This information can also be used for evaluation and assessment of criminal justice interventions.

Finally, subnational estimates provide a means for planning and allocating scarce government resources is an important concern to all stakeholders. Having the means to provide an effective and efficient allocation of resources targeted at localized crime problems remains a critical area for development.

To address the interest in small area estimates and to increase the value and utility of the general NCVS collection, the Bureau of Justice Statistics (BJS) has developed a small area estimation program that explores both direct and indirect estimation procedures. The specific details of many of these strategies are outlined in Cantor et al. (2010).

Among the research efforts is to investigate small area approaches to produce model-based estimates of crime at the state and possibly sub-state level. This paper summarizes

our progress to develop small area estimates for crime by state. We first review the findings of our previous empirical work to develop appropriate models for this problem (Li, Diallo, and Fay, 2012; Fay and Diallo, 2012). The third section presents the methodological approach, beginning with a summary of the small area model developed by Rao and Yu (1992, 1994). It also summarizes a modified version of this model termed the dynamic model (Fay and Diallo, 2012) and introduces the multivariate extension of this model. A fourth section notes implementation issues and reports preliminary results. The discussion section comments on the potential range of application of the new approach and compares it to related small area methods.

2. Summary of Previous Empirical Research

The development of a successful small area model typically requires identification of useful auxiliary variables, formation of a suitable model, and application of appropriate theory. The work reported here builds on exploratory work that was reported in previous papers. For completeness, we summarize the earlier findings directly supporting the current set of estimates.

2.1 Auxiliary Variables

The NCVS, begun in 1972 as the National Crime Survey (NCS), was predated by the Uniform Crime Report (UCR) of the Federal Bureau of Investigation, which originated in 1930 (Barnett-Ryan, 2007). The UCR was originally based on data aggregated by type of crime and reported by law enforcement jurisdictions. In the late 1980s, the UCR program was expanded and enhanced to allow for an incident- rather than an aggregate-based reporting, a system referred to as The National Incident Based Reporting System (NIBRS). Since the early 1990s, NIBRS has expanded and is currently used in a select set of jurisdictions, reporting detailed characteristics of individual crime incidents electronically. NIBRS data contribute directly to the UCR for participating jurisdictions. By design, the UCR can only reflect crimes reported to and recorded by the police; one of the original rationales for the creation of the NCS was to measure the impact of crime, including unreported crime, independently of the UCR.

Lynch and Addington (2007b) remark that the NCVS and UCR each have had their own advocates. The UCR can be affected by varying standards of reporting, and historically there have been occasional instances of deliberate efforts to distort the UCR statistics in particular jurisdictions. Besides the effect of unreported crime on the UCR, there are differences between the NCVS and UCR in their universe, criteria for classifying crime, and handling of complex crimes. Although most crimes in the NCVS have a corresponding version in the UCR Part 1 offenses, the largest component of violent crime in the NCVS, simple assault, lacks a corresponding version in the UCR Part 1 offenses.

An earlier but related goal of the overall project was to identify useful stratification variables for the NCVS redesign. (That effort, like much of the research reported here, used the Census Bureau's internal NCVS files, which includes geographic identifiers unavailable on the public files.) An investigation of the relationship between NCVS and UCR rates at the county level (Fay and Li, 2011) uncovered patterns that reappeared when a similar analysis was later performed at the state level (Li, Diallo, and Fay, 2012). In each case, the long-term average of the NCVS county or state crime rates for each type of crime was best predicted by a single variable from the UCR, as shown in Table 1. In most instances, the statistical evidence corresponds to conceptual expectations, that is, where the NCVS crime rate is best predicted by the corresponding rate from the UCR.

There are two exceptions: UCR forcible rape was found to be a far better predictor of NCVS aggravated assault than UCR aggravated assault. Secondly, it is also somewhat surprising that UCR forcible rape is also moderately successful at predicting simple assault, again displacing UCR aggravated assault as a predictor.

Table 1: Best UCR Predictor for Types of NCVS Crime at the State Level

NCVS Rate	*Best UCR Predictor*
Rape/sexual assault	Forcible rape
Robbery	Robbery
Aggravated assault	Forcible rape
Simple assault	Forcible rape
Household burglary	Burglary
Motor Vehicle Theft	Motor Vehicle Theft
Theft	Larceny

Source: Li, Diallo, and Fay (2012, Table 3)

2.2 Modeling Strategy

Although the UCR statistics do not precisely correspond to the unknown expected values of the NCVS, they appear to be the best available proxy for how the underlying NCVS crime rates might vary geographically and across time. Li, Diallo, and Fay (2012) presented various graphical summaries of the UCR data. When state values were plotted on a map of the U.S., different types of crime exhibited clearly different geographic patterns, suggesting that modeling crime by type of crime would be more effective and informative than simply modeling total crime. Secondly, comparison of UCR rates from a recent 3-year period to the corresponding period ten years earlier exhibited a high geographic correlation, suggesting substantial stability in the geographic pattern of crime rates over time.

Given this evidence, Li, Diallo, and Fay (2012) suggested fitting the Rao-Yu (1992, 1994) model to the data. This model will be described in more detail in the next section. To take advantage of the stability of the crime rates across time, the model employs information from the NCVS sample values in neighboring years rather than just for the target year being estimated. The model allows the NCVS sample estimates to be affected by sampling correlations over time, an important feature in this case because of the panel design and correlation over time arising from the first-stage selection of primary sampling units for NCVS, both of which induce sampling correlations.

3. The Multivariate Dynamic Model

3.1 The Rao-Yu Model

The Rao-Yu (1992, 1994) was summarized by Rao (2003). It builds on a linear mixed model for the population values, θ_{it}, t=1,…,T,

$$\theta_{it} = x_{it}'\beta + v_i + u_{it}$$

where

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

Furthermore, the u_{it}´s are assumed to form a stationary time series. The model for the observed sample values, y_{it}, is

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

where

e_{it} is random sampling error for area i, time t, with
$\mathbf{e_i} = (e_{i1}, \ldots, e_{iT})' \sim N_T(0, \mathbf{\Sigma}_i)$, and where
v_i, $\epsilon_{it,}$ and $\mathbf{e_i}$ are mutually independent.

As previously noted, $\mathbf{\Sigma}_i$ need not be diagonal, that is, the model can accommodate sampling covariances across time.

Rao and Yu (1992, 1994) derived the best linear unbiased predictor (BLUP) for this model for known values of the variance parameters σ_v^2, σ^2, and ρ. Adapting a method of moments approach developed by economists to estimate these parameters, they proposed a resulting empirical best linear unbiased predictor (EBLUP), although they noted considerable difficulty in applying the method because of the instability in estimating ρ. They also investigated an EBLUP based on estimating σ_v^2 and σ^2 given a presumed value of ρ. Rao (2003) summarized these results, again noting practical difficulties with the estimation of the variance parameters.

3.2 The Dynamic Model

Fay and Diallo (2012) noted that although the Rao-Yu model might provide a reasonable summary of the UCR crime data, the stationarity assumption appeared somewhat questionable in this application. They proposed instead a minor modification to the Rao-Yu model that would remove the stationarity requirement by modifying the random effect terms. The mixed model for the population values is

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

where

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

The sampling model is modified similarly

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

where again

e_{it} is random sampling error for area i, time t, with

$\mathbf{e_i} = (e_{i1}, \ldots, e_{iT})' \sim N_T(0, \mathbf{\Sigma}_i)$, and where

v_i, $\epsilon_{it,}$ and $\mathbf{e_i}$ are mutually independent.

Unlike the Rao-Yu model, ρ is not constrained. When $\rho > 1$, the model corresponds to a divergent situation in which areas become progressively more disparate. When $\sigma^2_{v*} = \sigma^2/(1-\rho^2)$ and $|\rho| < 1$, the dynamic model becomes equivalent a Rao-Yu model with $\sigma^2_v = 0$. But by dropping the stationarity assumption, the dynamic model is more appropriate for a situation in which the disparity among states dissipates over time.

Fay and Diallo (2012) compared the fit of the Rao-Yu model and the dynamic model to the UCR data at the state level, using a model that included only main effects for each year. For all of UCR variables, the dynamic model provided a better fit than the Rao-Yu model, in most cases by a statistically significant amount. Nonetheless, considering the size of the UCR data set, the improvements in fit were modest, suggesting that the Rao-Yu model would have provided an adequate alternative for the small area estimation application.

Fay and Diallo (2012) also reported successfully developing maximum likelihood (ML) and restricted maximum likelihood (REML) approaches to estimating the variance parameters, including ρ, both for the dynamic model and for the original Rao-Yu model, simply by following the general approach summarized by Rao (2003, Section 6.2) for the general linear mixed model. The possible use of ML and REML in this context was likely previously discovered by other researchers; for example, the REML approach was used for a spatial-temporal model that generalizes the Rao-Yu model (Marhuenda, Molina, and Morales, 2013).

3.3 The Multivariate Dynamic Model

Development of a multivariate version of the dynamic model was motivated by specific goals. A multivariate approach appeared to be the best solution to jointly model the components of crime and their sum, for example, burglary, motor vehicle theft, and other theft as the components of total property crime. The multivariate approach resolves the problem that univariate modeling of each component and their sum separately would produce a set of inconsistent estimates. The BLUP estimator is, by definition, a linear function of the observed **y,** and the results for BLUP for the general linear mixed model provides simultaneously the BLUP for any linear combination of the fixed and random effects (Rao, 2003, Section 6.2.1). As a consequence, the BLUP for the sum of crime rates by type of crime is the sum of the BLUPs for the components. (In our application, the result applies equally to 3-year averages of crime rates.) Rao (2003, section 8.1) reviewed a number of earlier applications of multivariate models to small area estimation, although they remain less frequently used than univariate models.

We implemented a multivariate version of the dynamic model. The population values for area i, time t can be represented as a vector $\boldsymbol{\theta}_{it} = (\theta_{it1}, \theta_{it2} \ldots)'$. By letting k index the components of the multivariate vector, a model for the population values can be expressed

$$\theta_{itk} = x'_{itk}\beta_k + \rho^{t-1}v^*_{ik} + u^*_{itk}$$

where

$\mathbf{v}_i^* = (\mathrm{v}_{i1}^*, \mathrm{v}_{i2}^*, \ldots)' \sim^{iid} N(0, \Sigma_{v*})$ is a vector of random effects for area *i* at *t*=1,

$\mathrm{u}_{\mathrm{i}1k}^* = 0$

$\mathrm{u}_{\mathrm{i}tk}^* = \rho \mathrm{u}_{\mathrm{i},t-1,k}^* + \epsilon_{\mathrm{i}tk}, \quad$ for $\mathrm{t} > 1$, where

$\boldsymbol{\epsilon}_{\mathrm{it}} = (\epsilon_{\mathrm{it}1}, \epsilon_{\mathrm{it}2} \ldots)' \sim^{\mathrm{iid}} N(0, \Sigma)$

and where Σ_{v*} and Σ are related to each other by $\Sigma_{v*(k,k')} = \boldsymbol{\sigma}_{v*k}^2 \mathbf{R}_{(k,k')} \boldsymbol{\sigma}_{v*k'}^2$ and $\Sigma_{(k,k')} = \boldsymbol{\sigma}_k^2 \mathbf{R}_{(k,k')} \boldsymbol{\sigma}_{k'}^2$, for a common correlation matrix, **R;** $\boldsymbol{\sigma}_{v*}^2 = (\sigma_{v*1}^2, \sigma_{v*2}^2, \ldots)'$; and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2 \ldots)'$. Note that the proposed model posits a single parameter ρ. The sampling model is

$$\theta_{itk} = \mathrm{x}_{\mathrm{it}k}' \beta_k + \rho^{\mathrm{t}-1} \mathrm{v}_{\mathrm{i}k}^* + \mathrm{u}_{\mathrm{it}k}^* + e_{itk}$$

where $\mathbf{v}_i$, $\boldsymbol{\epsilon}_{it,}$ and $\boldsymbol{e_i} = (e_{i11,} e_{i21}, \ldots, e_{iT1}, e_{i21,} \ldots)'$ are mutually independent.

Again, existing theory (Rao, 2003, section 6.2) provides the general mathematical results to implement this model, including ML and REML estimation of the parameters and mean square error estimation for the REML results based on extensions of the methods begun by Prasad and Rao (1990). We have developed functions in R (R Core Development Team, 2013) to implement the calculations. (The functions for ML and REML produce the univariate version of the dynamic model as a special case.)

3.4 Monte Carlo Comparison

Simple Monte Carlo comparisons illustrate how the multivariate version of the dynamic model can potentially improve upon the univariate version. Two simulations were performed, each based on 5,000 generated data sets. The two simulations shared a number of features in common. Each was based on 32 areas with observations on two correlated variables over 10 years. The parameters were chosen to satisfy simultaneously the conditions of both the dynamic and Rao-Yu models. A high value, ρ = 0.95, was selected to illustrate situations with high gains from the time-series aspects.

In the first simulation, the two variables were identically distributed, with both components of the vector were assigned $\sigma_k^2 = 0.9\ (1 - \rho^2)$ and $\sigma_{\mathrm{v}k}^2 = 0$, in terms of the Rao-Yu model. Alternatively, the parameters were $\sigma_k^2 = 0.9\ (1 - \rho^2)$ and $\sigma_{\mathrm{v}*\mathrm{k}}^2 = 0.9$ in terms of the dynamic model. The correlation between the two values of $\epsilon_{\mathrm{it}1}$ and $\epsilon_{\mathrm{it}2}$ was set to 2/3, which was also the correlation between v_{i1}^* and v_{i2}^*. The 64 sampling errors were drawn from the standard normal.

The second simulation was similar to the first, defining the first variable in the same way as the first simulation but modifying the second variable by rescaling the problem. The expected value of the second variable was 9 times the expected value for the first variable, and the variance components, both the variances of the random effects and the sampling variances, were multiplied by a factor of 9. Correlations and other features of the first simulation were otherwise preserved. Table 2 presents the results of the simulations. Because the same random seeds were used to start both simulations, the results for the first component were identical.

Table 2: Comparison of Univariate and Multivariate Dynamic Models for Simulated Data Sets Generated Under the Rao-Yu Model

Average MSE or estimated MSE	*Simulation 1*	*Simulation 2*
First component:		
MSE of univariate SAE	.267	.267
Estimated MSE for univariate SAE	.273	.273
MSE of multivariate SAE	.249	.249
Estimated MSE of multivariate SAE	.254	.254
Second component:		
MSE of univariate SAE	.268	2.414
Estimated MSE for univariate SAE	.272	2.451
MSE of multivariate SAE	.250	2.246
Estimated MSE of multivariate SAE	.254	2.281
Sum of two components:		
MSE of univariate SAE	.677	3.108
MSE of multivariate SAE	.655	2.965
Estimated MSE of multivariate SAE	.656	2.981

In both simulations, information from one component improves the estimation of the other by about 7%. The sum is improved by about 3% in the first simulation and by almost 5% in the second. The multivariate approach thus improves on the univariate approach in all cases. And in all cases the estimated MSEs on average closely predict the actual average MSEs, over-estimating by slight amounts.

4. Implementation and Results

4.1 Implementation

Using the multivariate dynamic model fitted with REML as our basic approach, we have produced state-level small area estimates of crime for the period 1997-2011. The results have been expressed in the form of 3-year averages of rates, that is, 13 sets of state estimates corresponding to the overlapping periods 1997-1999, 1998-2000, etc., through 2009-2011. Showing results as 3-year averages smooths the findings and anticipates possible future release of direct 3-year averages for large states from an expanded NCVS.

Separate estimates of property crime were produced for burglary, motor vehicle theft, and other theft. In spite of high interest in rates for rape and sexual assault, we decided not to attempt to estimate this characteristic separately because the NCVS data on rape and sexual assault at the state level are quite sparse, clearly challenging the normal assumption that was the basis of the EBLUP theory. Instead, NCVS data on rape and sexual assault were combined with aggravated assault into a category that could be called *combined serious assault*. The small area estimates of violent crime are therefore disaggregated into combined serious assault, simple assault, and robbery.

We addressed a separate interest for statistics on the relationship of the victim to the perpetrator; in particular, the issue of violent crimes committed by intimate partners has been examined in a series of BJS publications (e.g., Truman and Planty, 2012; Catalano, 2012). Predominantly, but not exclusively, the victims of these crimes are women. We produced small area estimates of violent crimes by three categories of perpetrators: strangers, intimate partners, and all others. The 9-cell cross-classification of type of violent crime by category of perpetrator again yields data too sparse at the state level to

consider modeling with our approach. Instead, we modeled violent crime by type of crime and violent crime by perpetrator, separately.

The original plan was to fit three multivariate models, each with three components: (1) property crime by type, (2) violent crime by type, and (3) violent crime by perpetrator. The estimated totals for violent crime would therefore require reconciliation, hopefully by a small adjustment. In fact, this initial approach required modification when it became evident that the multivariate approach may be vulnerable to departures from normality arising from sparse data.

The NCVS estimates that motor vehicle theft is now less than 5% of total property crime and many small states have reported frequencies of zero in several years. In one state the number of reported instances was quite relatively large but still small in absolute terms; the effect on the multivariate model was to produce an implausible estimate of total property crime out of line with similar states and the rest of the sample evidence. Two multivariate models were then fitted, each with two categories. The first model, for total property crime, modeled burglary and all theft; the second, for theft, modeled motor vehicle theft with other theft. The estimates of the second model were proportionately adjusted to agree with the results from the first model for total theft. The first model consequently provided the estimates for total property crime.

Similarly, intimate partner violence has been typically less than 20% of total violent crime. Again, two models were fitted. The first model for total violent crime contrasted crimes by strangers with crimes by non-strangers. The second contrasted crimes by intimate partners with crimes by all other non-strangers. Again, the totals from the second model were proportionately adjusted to the results from non-strangers in the first model.

The model for violent crime by type was retained as a 3-category multivariate model because the least frequent of the three categories, robbery at approximately 10% of the total, did not fit conceptually into either of the remaining two categories. The results from this model were proportionately adjusted to the estimated total violent crime from the first of the two models for crimes by perpetrator.

The estimates described here reflect a change in the estimation procedure for series incidents, introduced by BJS in 2012 (Lauritsen et al. 2012). Until 2011, the annual publication of crime statistics had excluded repeated crimes where the respondent could not provide separate details; for 2011, BJS began to incorporate the number of such repeated incidents, trimming the number at 10 (Truman and Planty 2012). Revised estimates back to 1993 are available from the NCVS Victimization Analysis Tool (http://www.bjs.gov/index.cfm?ty=nvat). The revised estimator increased some estimates of violent crime appreciably, such as intimate partner violence, for example (Truman and Planty 2012). The revision considerably increases the influence of some observations, with the effect of making the sample estimates less normally distributed. In turn, the reduced normality makes the small area estimation problem more challenging.

A few small states lacked NCVS sample entirely, and estimates were produced for these states based on the fixed-effect regression model.

The sampling variances and covariances used in fitting the dynamic model were modeled, taking into account the large difference in observed design effects and correlations over time between large self-representing (SR) metropolitan areas (CBSAs)

and less-populous non-self-representing (NSR) areas in the design. The non-self-representing areas exhibited both larger average design effects and larger correlations across time. Rather than use the actual design, the variance model is based on a "public model" of the design that identified likely self-representing areas based on their population and the general design strategy for NCVS described in publicly available documents (Fay and Li, 2012). For example, the public model indicates that North Carolina would tend to have a larger proportion of the population covered by NSR primary sampling units, compared, for example, to California. Use of direct estimates of sampling variances and covariances was out of the question because of the high instability of the variance estimator at the state level. Modeling of sampling variance has been a feature in many previous applications, including Fay and Herriot (1979). Berg and Fuller (2012) presented recent work on the role of modeling sampling variances for use in SAE applications.

4.2 Findings

At the time of this writing, the estimates are under review, so we have withheld most of the results pending public release. A modified version of the estimates, prepared with the same methods but omitting the NCVS data for 2006, will be considered. The NCVS data for 2006 produced an anomalous one-year bump in crime rates (Rand, 2008; Rand and Catalano, 2007).

Figure 1 presents a comparison to illustrate that the time-series aspect of the model substantially preserves long-time averages in crime rates for the 36 largest states, which have a 2010 population of 2 million or more. Based on the public model of the design, all of these states may reasonably be expected to be included in the NCVS sample, which is in fact the case. Each plot compares the 15-year averages of the directly observed NCVS rates to an average formed by averaging the SAE estimates for 1997-1999, 2000-2002, 2003-2005, 2006-2008, and 2009-2012. For property crimes, the results are virtually equal, showing that the SAE estimates of property crime for these states mirror the long-term average from NCVS sample. Because violent crime is much less frequent than property crime and its coefficient of variation is consequently larger, the SAE estimates largely track the NCVS differences but also exhibit more regression toward the mean, such that the SAE estimates are somewhat less dispersed than the direct estimates.

When released the comparisons may also be shown in a table with individual states identified. The intent is to address possible skepticism from potential users of the data. The comparison illustrates that even though the estimates for individual years have been smoothed, the estimates over an extended period largely reflect the direct evidence from the NCVS.

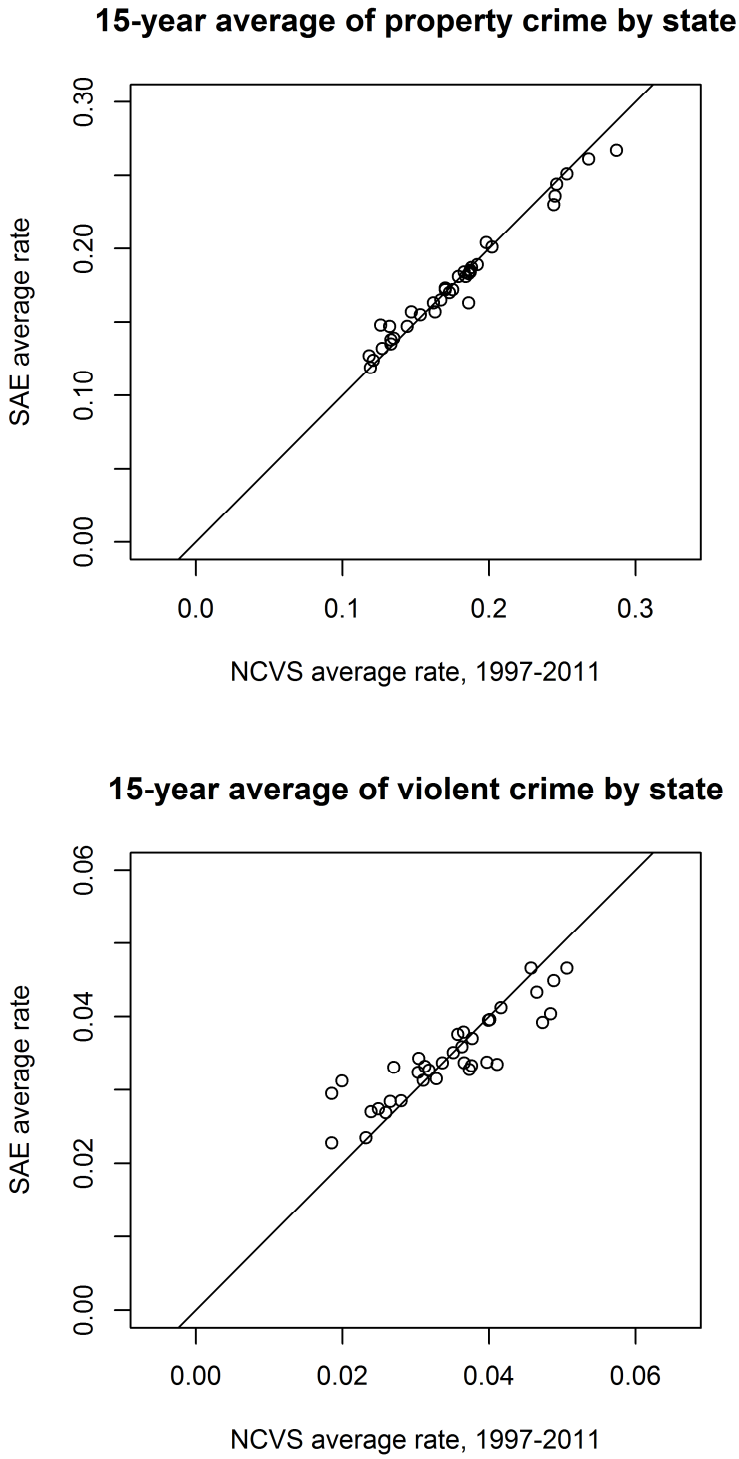

Figure 1: Comparisons of small area estimates with direct estimates when averaged over 15 years, for property crime and violent crime, in states of 2 million or more persons in 2010. A 45-degree line through the origin is shown on each plot. SAE averages for violent crime, generally above the line to the left and below the line to the right, exhibit regression toward the mean more than the SAE for property crime.

5. Discussion

In July, 2013, the NCVS sample design was expanded to permit the publication of direct estimates of 3-year averages of crime rates in the 11 largest states. The expected precision of the direct estimates in these states, measured in terms of sampling variance, will likely exceed the reliability, measured in terms of mean square error, currently achieved by the SAE described here. A further expansion to include direct estimates in more states is envisioned to begin in a few years with the redesign of the survey. Thus, the longer term vision for the NCVS includes expanded state information through direct estimation, but this future may evolve to include a continued role for SAE, such as to provide information for smaller states and for substate areas such as counties and cities.

Extending the results to large counties and cities is an attractive next step. Large counties, such as those of 1 million persons or more, are typically self-representing. A preliminary analysis suggests that the results of modeling this group of counties will produce useful estimates. New models for variances and covariances will be required, to take advantage of the potential sampling variance information available from SR areas, where variances may be estimated at the between-segment rather than between-PSU level. The experience from modeling states cautions against applying the multivariate model to highly sparse data. Because of small sample sizes, it may be necessary to implement a reduced set of models, possibly eliminating the multivariate approach when modeling counties or cities.

The application illustrates potential gains from a multivariate approach to the Rao-Yu and the related dynamic model. For example, in a recent publication of SAE estimates of cell phone use (Blumberg et al., 2012), the authors suggest the possibility of taking a multivariate approach to their application, which is based on the Rao-Yu model. Our findings might further encourage them to attempt to do so.

At the same time, we also find that unfavorable conditions, such as non-normal behavior arising from sparse data, can limit the usefulness of the multivariate approach based on EBLUP estimation. Additional research could identify the necessary conditions sufficient in practice to favor the application of the multivariate approach.

Acknowledgements

The results reflect the views of the authors and not necessarily those of Westat, the Bureau of Justice Statistics, or the Census Bureau. We wish to thank Meagan Wilson and her Census Bureau colleagues for resource support enabling this research. We also thank Graham Kalton for his comments on the paper.

References

Barnett-Ryan, C. (2007), "Introduction to the Uniform Crime Reporting Program," in Lynch, J.P. and Addington, L.A. (eds.) (2007), *Understanding Crime Statistics: Revisiting the Divergence of the NCVS and UCR*, pp. 55-89.

Berg, E.A. and Fuller, W.A. (2012), "Estimators of Error Covariance Matrices for Small Area Prediction, *Computational Statistics and Data Analysis*, 56, 2949–2962.

Blumberg, S.J., Luke, J.V., Ganesh, N., Davern, M.E., Boudreaux, M.H., and Soderberg, M.S. (2011), "Wireless Substitution: State-level Estimates from the National Health

Interview Survey, January 2007-June 2010," National Health Statistics Reports: No. 39. Hyattsville, MD: National Center for Health Statistics.

Catalano, S. (2012), "Intimate Partner Violence," NCJ 239203, Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics.

Cantor, D., Krentzke, T., Stukel, D., and Rizzo, L. (2010), "NCVS Task 4 Report: Summary of Options Relating to Local Area Estimation," issued by Westat to the Bureau of Justice Statistics, May 19, 2010.

Fay, R.E. and Diallo, M.S. (2012), "Small Area Estimation Alternatives for the National Crime Victimization Survey," *Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, VA, pp. 3742-3756.

Fay, R.E. and Herriot, R.A. (1979), "Estimates of Income for Small Places: An Application of James-Stein Estimation to Census Data," *Journal of the American Statistical Association*, 74, 269-277.

Fay, R.E. and Li, J. (2012) "Rethinking the NCVS: Subnational Goals through Direct Estimation," presented at the 2012 Federal Committee on Statistical Methodology Conference, Washington, DC, Jan. 10-12, 2012, available at http://www.fcsm.gov/12papers/Fay_2012FCSM_I-B.pdf.

Lauritsen, J., Owens, J., Planty, M., and Truman, J. (2012), "Methods for Counting High Frequency Repeat Victimizations in the National Crime Victimization Survey," NCJ 237308, Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics.

Lauritsen, J. and Schaum, R. (2005). "Crime and Victimization in the Three Largest Metropolitan Areas, 1980-1998." NCJ 208075, Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics.

Li, J., Diallo, M.S., and Fay, R.E., (2012) "Rethinking the NCVS: Small Area Approaches to Estimating Crime," presented at the 2012 Federal Committee on Statistical Methodology Conference, Washington, DC, Jan. 10-12, 2012, available at http://www.fcsm.gov/12papers/Li_2012FCSM_I-B.pdf.

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

Lynch, J.P. and Addington, L.A. (2007b), "Introduction," in Lynch, J.P. and Addington, L.A. (eds.) (2007), *Understanding Crime Statistics: Revisiting the Divergence of the NCVS and UCR*, pp. 3-13.

Marhuenda, Y., Molina, I., and Morales, D. (2013), "Small Area Estimation with spatio-temporal Fay-Herriot Models," *Computational Statistics and Data Analysis*, 58, 308-325.

Prasad, N.G.N., and Rao, J.N.K. (1990), "The Estimation of Mean Squared Error of Small-Area Estimators," *Journal of the American Statistical Association*, 85, 163-171.

R Core Team (2013), "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rao, J.N.K. (2003), *Small Area Estimation*, John Wiley & Sons, Hoboken, NJ.

Rao, J.N.K. and Yu, M. (1992), "Small Area Estimation Combining Time Series and Cross-Sectional Data," *Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, VA, pp. 1-9.

_____ (1994), "Small Area Estimation by Combining Time Series and Cross-Sectional Data," *Canadian Journal of Statistics*, 22, 511-528.

Rand, M. (2008), "Criminal Victimization: 2007," NCJ 224390, BJS website, Dec. 2008.

Rand, M. and Catalano, S. (2007), "Criminal Victimization: 2006," NCJ 219413, BJS website, Dec. 2007.

Skogan, W. (1977). “Dimensions of the Dark Figure of Unreported Crime,” *Crime & Delinquency, 23(1):41-50.*

Truman, J.L. and Planty, M. (2012), “Criminal Victimization: 2011,” NCJ 239437, BJS website, Oct. 2012.