Comparing Generalized Variance Functions to Direct Variance Estimation for the National Crime Victimization Survey

Bonnie Shook-Sa, David Heller, Rick Williams, G. Lance Couzens, and Marcus Berzofsky

RTI International
3040 Cornwallis Rd, Research Triangle Park, NC 27709

Abstract

Currently, the National Crime Victimization Survey (NCVS) relies on generalized variance functions (GVFs) for the calculation of standard errors and for significance testing. However, GVFs developed for the NCVS are cumbersome when multiple estimates are produced, do not allow for complex analyses such as regression modeling, and the accuracy of GVF estimates for outcomes not included in developing the GVF parameters is unknown. Use of GVFs requires knowledge about the correct GVF parameters and formulas to use, and these decisions are dependent on the outcome of interest.

Direct variance estimation techniques such as Taylor Series Linearization (TSL) and Balanced Repeated Replication (BRR) allow variances to be calculated using existing software packages, making estimation more straight forward for most users. Both estimation techniques require study design data (i.e. stratification variables and primary sampling units) in either the creation of the weights (BRR) or in the variance estimation itself (TSL), so resulting estimates accurately reflect the complex survey design. While the NCVS public use file contains some design variables, the full set of variables are not publically available due to disclosure concerns. This paper presents the first evaluation of the feasibility of direct variance estimates based on the available design variables and addresses logistical challenges imposed by direct estimation techniques, specifically those encountered when estimating victimization rates based on multiple input files and sampling weights.

We discuss the complexities associated with calculating direct variance estimates for the NCVS and compare direct variance estimates (TSL and BRR) to estimates produced using GVFs. We evaluate these methods for multiple outcome types (e.g. totals and rates), subgroups of interest (e.g. gender, race, and age), and for single and multi-year estimates. Additionally, we develop recommendations for users of the NCVS public use files regarding NCVS variance estimation.

1. Introduction

The National Crime Victimization Survey (NCVS), sponsored by the Bureau of Justice Statistics (BJS), provides estimates of the incidence and characteristics of criminal victimization in the United States. When calculating NCVS estimates, researchers must take into account the complex stratified, four-stage sample design and analysis weights. Stratification, clustering, and variation in analysis weights all affect the variances of survey parameters, and not appropriately accounting for these factors during estimation can lead to invalid results (Cochran, 1977).

Two broad methods exist for calculating variances of estimates from complex sample designs: Generalized Variance Functions (GVFs) and direct variance estimation. GVFs model the design-consistent variances for multiple survey estimates to obtain GVF parameters. Using the formulas and parameters from the GVF models, users are able to calculate approximations of variances without knowledge of the sample design. Direct variance estimation uses software that accounts for complex sample designs. Two direct variance techniques are Taylor Series Linearization (TSL) and Balanced Repeated Replication (BRR).

Currently, BJS uses GVFs to calculate variances of NCVS estimates. However, the GVFs developed for the NCVS do not allow for complex analyses such as regression modeling, are cumbersome when multiple estimates are produced, and produce GVF estimates for outcomes not included in developing the GVF parameters that are of unknown accuracy. Use of GVFs requires knowledge about the correct GVF parameters and formulas to use, and these decisions are dependent on the outcome of interest.

Direct variance estimation has not been used for the NCVS because two analysis files and two weights are needed for the calculation of key NCVS estimates (victimization rates): a population weight from either the household or person-level file and a victimization weight from the incident file. The population weight represents the number of persons or households in a domain of interest. The victimization weight represents the number of victimizations experienced by the person or household. In order to properly calculate the variance of a rate both weights are required. However, currently, no software package allows for two weight values to be used in the calculation of the variance, making it difficult to use direct variance estimation.

This paper examines the feasibility of using direct variance estimation for the NCVS. It compares GVF estimates to two direct variance estimation methods (TSL and BRR). When comparing direct variance estimation to the current GVF approach, the following areas are addressed:

1. Single year estimation
2. Pooled year estimation
3. Cross single year estimation
4. Cross pooled year estimation

2. Variance Estimation Options

The NCVS sample consists of approximately 50,000 sample housing units selected each year with a stratified, multi-stage cluster design. The Primary Sampling Units (PSUs) composing the first stage of the sample include counties, groups of counties, or large metropolitan areas. PSUs are further grouped into strata. Large PSUs are included in the sample automatically and each is assigned its own stratum. These PSUs are considered to be self-representing (SR) since all of them are selected. The remaining PSUs, called non-self-representing (NSR) because only a subset of them is selected, are combined into strata by grouping PSUs with similar geographic and demographic characteristics, as determined by the decennial Census used to design the sample. A single NSR PSU is selected from each stratum. For analytic purposes, the SR PSUs are each separated into two pseudo-PSUs and labeled as coming from the same pseudo-stratum. Each NSR PSU is paired with a second NSR PSU selected from a similar stratum and labeled as two pseudo-PSUs coming from the same pseudo-stratum. The pseudo-PSUs and pseudo-strata are important concepts for the variance estimation methods described below and are used to describe the sample design when analyzing the data.

The NCVS sample of PSUs is drawn every 10 years from the decennial Census and used until the next decennial Census is available at which point a new sample of PSUs is selected. At approximately mid-decade, sample selection from the most recent Census is phased in, and prior to that, sample selection is based on the Census before the most recent one. For example, prior to 1995, the sample was drawn from the 1980 decennial Census. From January, 1995 until December, 1997, the sample drawn from the 1990 Census was phased in. From January, 1998 until approximately 2005, the complete NCVS sample was drawn from the 1990 Census. From 2005 through 2007, samples from the 2000 Census were phased in. As will be shown, the transition between decennial PSU samples is important when implementing direct variance estimation.

Because of the continuing nature of the NCVS, a rotation scheme is used to avoid interviewing the same household indefinitely. A sample of housing units is divided into six rotation groups, and each group is interviewed every six months for a period of three years. Within each of the six rotation groups, six panels are designated. A different panel is interviewed each month during the six-month period. Within each selected NCVS household, all persons aged 12 and over are eligible to complete the interview.

Multistage sample designs like the one employed in the NCVS complicate data analysis since the individual person and household observations are not independent (Wolter, 1985). The observations are correlated due to having been selected from geographic or household clusters of likely similar survey units (housing units within a PSU and persons within a household are likely correlated). Also, using the same sample of PSUs for a ten-year period, combined with repeated interviews of the same housing units over rotating three year periods, causes estimates from years using the same PSU sample to be correlated.

In the sections that follow, three methods for variance estimation are discussed and compared. The first is the use of generalized variance functions (GVFs), which have been available for use with the NCVS public use data since its inception in 1992. The other two, Taylor series linearization (TSL) and balanced repeated replication (BRR), are two direct variance estimation methods that are being explored as alternative methods for use with the NCVS public use data.

Direct variance estimation methods use statistical software designed to calculate the variance of an estimate directly from the full dataset. In order to implement direct variance estimation, users must organize and code the data so that each observation is associated with the stratum and PSU from which it was selected. To this end, the public use data files include the following two variables:

> **Pseudo-stratum**: The variable designating the pseudo-stratum code associated with each observation is created from the sampling strata used to select the PSUs.
>
> **Half-sample**: The variable designating the pseudo-PSU code associated with each observation is created from the sampling PSUs selected into the sample. The term "half-sample" is used since there are two pseudo-PSUs from each pseudo-stratum which approximately divide the sample in half.

The terms "stratum" and "PSU" will be used throughout this paper to refer to the variables pseudo-stratum and half-sample.

Exhibit 1 presents the number of strata included on the NCVS public use files from 1993 through 2010 with each stratum containing two PSUs. The exhibit also presents the grouping of years for which Decennial Census data were used to select the sample of PSUs contributing to the data for the years in each group. Except for issues arising from the phase-in/phase-out periods, the PSUs used to select the data within a Year Group are the same for each year, whereas for the between Year Groups the samples of PSUs are different. Thus, the data between Year Groups are assumed to be independent but the data within a Year Group are assumed to be cluster correlated within the PSUs across years. These assumptions will be used for direct variance estimation. Although these assumptions are only approximately true due the phase-in/phase-out process, the assumptions are necessary since the public use data files do not contain the level of detail needed to separately account for the overlap of PSUs during the phase-in/phase-out period. The approximations will, however, support appropriate direct variance estimation.

Exhibit 1. Grouping of Years by Decennial Census and Number of Strata by Year

Grouping of Years by Decennial Census	Year	Number of Strata
Year Group 1 PSU sample primarily from the 1980 Decennial Census	1993	164
	1994	164
	1995	164
	1996	164
Year Group 2 PSU sample primarily from the 1990 Decennial Census	1997	143
	1998	143
	1999	143
	2000	143
	2001	143
	2002	143
	2003	143
	2004	143
	2005	144
Year Group 3 PSU sample primarily from the 2000 Decennial Census	2006	160
	2007	160
	2008	160
	2009	160
	2010	160

2.1 Generalized Variance Functions

Within the NCVS, GVFs are estimated by the U.S. Census Bureau and approximate the variance of an estimate as a function of readily available information about the estimate. The process starts by selecting a set of NCVS estimates and calculating their associated variances. Over the years, the Census Bureau has estimated the variances using different direct variance estimation methods, including TSL, jackknife, BRR, and successive difference replication. The first three methods are widely used (Wolter, 1985), but the latter is a more specialized method described in Fay and Train (1995) and Ash (2010). Modeling methods, like those described in Wolter (1985, Chapter 5), are then used to model the variance as a function of such values as the estimate, the sample size or the population size, or other characteristics related to the sample design (such as location or urban vs. rural) or to the respondent (such as age, race, or marital status). It is also common that separate models are required for various types of estimates, for example, victimization rates, totals, or percentages. The resulting models are called generalized variance functions, or GVFs.

Although GVFs have the advantage of allowing users to calculate design-consistent variance estimates without knowledge of the sample design, they are limited to the specific situations for which they are designed. For example, when studying the relative victimization rate of African American versus White Americans, GVFs are available for the two separate victimization rates, but not for the relative victimization rate (or, the ratio of the two individual victimization rates). Moreover, separate GVFs are needed for different victimization types and for each year. Thus, when conducting a large analysis spanning several years and victimization types, many different GVFs are needed, which makes it difficult to manage the analysis.

Importantly, reporting crime victimization statistics that either exclude or include series or repeat victimizations, is a complicating factor for this analysis. Series victimization reporting is allowed when a respondent is unable to separate the facts of six or more similar victimizations occurring within a six month period. In cases like these, the respondent can report the number of victimizations and only the details of the most recent event. Until recently, BJS reported crime statistics *excluding* series reported victimizations, but BJS now reports crime victimization statistics *including* series victimizations (Lauritsen, Owens, Planty, Rand, & Truman, 2012). Up until this change, the U.S. Census Bureau created GVFs for estimates excluding series victimizations. In July 2013, the Census issued updated GVFs for estimates including series victimizations for years 2008 through 2012. Although most sections of this paper will use data including series victimizations, some sections will use data excluding series victimizations for comparison to past work or situations in which GVFs for estimates including series victimizations are not available. Each situation will be identified clearly.

2.2 Taylor Series Linearization

For a stratified multistage cluster sample like the one used for the NCVS, there is an unbiased variance estimator for a linear statistic. An example of a linear statistic is the estimated total number of victimizations for a year given by $Y = \sum_{j=1}^{n} w_j y_j$ where w_j and y_j are the analysis weight and the number of victimizations incurred by the j^{th} participant in the survey, respectively. The variance estimator is based on the commonly used assumption that the PSUs in a multistage sample were selected with replacement. Although replacement PSU selection is almost never done, it is a good approximating assumption when the sampling fraction (i.e., the ratio of the number of PSUs selected and the total number of PSUs in the stratum) among the PSUs is small. For the NCVS, there are only two PSUs per stratum selected out of a large number of PSUs available per stratum, so the replacement PSU sampling assumption is appropriate. The variance estimation formula is

$$V_{TSL}(Y) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (Y_{hi} - \bar{Y}_h)^2$$

where $Y_{hi} = \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}$ and $\bar{Y}_h = \sum_{i=1}^{n_h} y_{hi} / n_h$. The subscripts have been expanded to include strata (h), PSUs (i), and respondents (j), and with n_h being the number of PSUs in a stratum and m_{hi} the number of respondents in a PSU. This variance estimator has been shown to be unbiased for linear statistics (Särndal, Swensson, and Wretman, 1992; Williams, 2000).

When considering a nonlinear statistic, the TSL method replaces the nonlinear statistic with a first order Taylor series linear approximation and then uses the above variance estimator with the linear approximation data to estimate the variance of the nonlinear statistic. The resulting variance estimate is a consistent estimate of the variance of the nonlinear statistic. For example, the victimization rate is estimated by $R = Y/X$ where Y is the estimated total number of victimizations as just described and $X = \sum_{j=1}^{n} w_j$ is the estimated total number of people (for personal crimes) or households (for property crimes) in the population. Following the descriptions in Wolter (1985, Section 6.5) or Williams (2008), it can be shown that the linearized values for a ratio are $z_j = (y_j - Rx_j)/X$.

The TSL method is widely implemented in statistical analysis software packages, such as SUDAAN, SAS, Stata, and SPSS. All of these analysis packages automatically determine the linearized values for a wide range of statistics without the need for user input. However, the analysis packages require the user to specify the strata and PSUs used to select the sample so that the variance can be estimated appropriately. For an estimate based upon data from a single year, the variables Pseudo-stratum and Half-sample are the variables that specify the strata and PSUs to the analysis package. The situation is slightly more complex when analyzing data across years because of the use of the same PSUs across 10-year intervals and the repeated interviewing of the same households over three years. In this situation, the same strata and PSUs are used across years within the Year Groups shown in ***Exhibit 1***. The key is to group data across the years by the strata and PSUs used to select the data. Thus, ***Exhibit 2*** illustrates how to create cross-year strata so that data within the same Year Group use the same strata and PSUs in the variance calculation, which will capture the statistical correlation among these data. On the other hand, the cross-year strata will separate the data from two different Year Groups in the variance calculation and treat the different Year Groups as statistically independent.

Exhibit 2. Cross Year Strata and PSUs

Cross-Year Strata		PSUs	
Year Group	**Pseudo-stratum (V2117)**	**Half-sample (V2118)**	**Years of Data**
1	1	1	1993–1996
1	1	2	1993–1996
1	2	1	1993–1996
1	2	2	1993–1996
⋮	⋮	⋮	⋮
1	164	1	1993–1996
1	164	2	1993–1996
2	1	1	1997–2005
2	1	2	1997–2005
2	2	1	1997–2005
2	2	2	1997–2005
⋮	⋮	⋮	⋮
2	144	1	1997–2005
2	144	2	1997–2005
3	1	1	2006–2010
3	1	2	2006–2010
3	2	1	2006–2010
3	2	2	2006–2010
⋮	⋮	⋮	⋮
3	160	1	2006–2010
3	160	2	2006–2010

2.3 Balanced Repeated Replication

BRR is another commonly used direct variance estimation method for complex sample surveys (Lumley, 2008). Like the TSL method, BRR takes advantage of the with replacement sampling assumption of the PSU sample. BRR is most easily implemented for a stratified sample with 2 PSUs selected per stratum like the pseudo-strata and pseudo-PSUs of the NCVS. The method proceeds by separating the NCVS into half-samples created by selecting one PSU from each stratum and the weights of observations in the selected half-sample are doubled, while the weights for the remaining observations are set to zero. A half-sample estimate of a statistic (victimization total, rate, or percent) is then obtained from the half-sample data. A large number of half-samples are generated along with a corresponding set of half-sample estimates denoted as $\theta_1, \ldots, \theta_G$ where G is the total number of half-samples created. The variance is then estimated by

$$V(\theta) = \sum_{g=1}^{G} \left(\theta_g - \theta\right)^2 / G$$

where θ is the estimated statistic from the full NCVS sample. The set of half-samples is usually selected so that they are in full orthogonal balance, in which case an efficient and consistent estimate of the variance in obtained. The conditions and methods for creating half-samples with full orthogonal balance are described by Wolter (1985, Chapter 3).

Similar to the TSL method, special consideration is needed to account for the overlap in strata and PSUs within a Year Group. The same cross-year strata and PSUs presented in ***Exhibit 2*** can be used when forming the BRR half-samples. When analyzing data from a single Year Group, the strata and PSUs specific to that Year Group are used to form the half-samples for BRR estimation. Once formed, the same half-samples are used for all years within the Year Group. For example, for Year Group 1, there are 164 strata each with 2 PSUs for all the years of data in Year Group 1 and the half-samples would be formed from these strata and PSUs. For analyses using data from two Year Groups, half-samples are needed using the strata and PSUs from both Year Groups. For example, if data were being compared across Year Groups 1 and 2, say pooled data from 1993–1996 compared to 1997–1999, then half-samples would be created from the combined 208 (164 + 144 = 208) strata from Year Groups 1 and 2. Finally, if all 3 Year Groups were included in the analysis, half-samples would be created from all 368 strata (164 + 144 + 160 = 368). In any of these cases, the data within a Year Group would be included or excluded from the same half-samples so as to capture the correlations due to sharing the same PSUs in a Year Group.

3. Preparing NCVS Data Files for Direct Variance Estimation

Three NCVS data files are needed for NCVS estimation: the household-level file, the person-level file, and the incident-level file. The household-level file contains one record for each sampled household in the NCVS per reporting period. It contains data from the household screening interview, which assesses whether a household experienced any property crimes during the previous six months. The household-level weight is contained on the household file, and is used to calculate household population estimates needed for the denominators of property victimization rates.

The person-level file contains data for each household member aged 12 or older in responding NCVS households. Each record corresponds to a sampled person within a reporting period. Data come from the personal screening interviews which are administered to all eligible and participating household members. The screening interview determines whether a person experienced a personal victimization during the previous six months. The person-level weight, contained on the person file, is used to calculate population estimates used for the denominators of personal victimization rates.

In most cases, the incident-level file contains one record for each victimization reported by NCVS respondents. It contains both property crimes reported by the household respondent (i.e., household burglary, motor vehicle theft, and theft) and personal crimes reported by any NCVS respondent (i.e., rape/sexual assault, robbery, aggravated assault, simple assault, and personal theft). The incident file contains data to classify victimizations based on crime type as well as details of each victimization drawn from the incident report (e.g. persons present, victim-offender relationship, weapon use). If the respondent reports six or more criminal incidents of a similar nature but cannot

recall specific details of each incident, the incidents are collapsed into a single record on the incident-level file and the total victimization count is recorded. These types of victimizations are called series victimizations. The victimization weight is contained on the incident-level file and is used to estimate the number of criminal victimizations with a given characteristic. It is used to estimate victimization totals and proportions and to estimate the numerators of personal and property victimization rates.

Victimization totals and proportions are calculated from a single file using a single weight (incident-level file and victimization weight, respectively). Therefore, the only steps needed to prepare for direct variance estimation of victimization totals and proportions are to: 1) create the year group variable as discussed in ***Section 2***, and 2) to ensure that all strata and PSUs are represented on the incident-level file. Because the incident-level file only contains data for persons and households reporting victimizations, PSUs where no respondents reported victimizations are not represented. To ensure that the NCVS design is appropriately represented, dummy records should be added to the incident-level file for any PSUs not represented. Following these steps, the incident file is ready for direct variance estimation of victimization totals and proportions.

Calculating victimization rates requires knowledge about the total population and the victimized population. Victimization rates are calculated by taking the ratio of the number of victimizations to the total population and multiplying this ratio by 1,000. The numerator of the victimization rate is estimated from the incident-level file, using the victimization weight. For property crimes, the denominator is calculated from the household-level file with the household weight. For personal crimes, the denominator is calculated from the person-level file with the person weight. Because estimates of victimization rates are based on two files and two sets of weights, which current software packages cannot accommodate, pre-processing is needed prior to calculating direct variance estimates. Victimization summaries, unweighted counts of victimizations with the characteristic(s) of interest, must be calculated from the incident-level file and moved to the person and household files prior to direct variance estimation. Furthermore, the victimization weights must be parsed out into their components and applied to estimates, as appropriate. These pre-processing steps are outlined in detail in the NCVS direct variance user's guide (Shook-Sa, Couzens, & Berzofsky, in press), which will be made available to NCVS analysts.

4. Single Year Estimates

This section explores single year victimization rate and total estimates and compares the GVF, TSL, and BRR variance estimation approaches. The following victimization types are included:

Personal Victimization Types
- Rape/sexual assault
- Robbery
- Aggravated Assault
- Simple Assault
- Personal theft

Property Victimization Types
- Household burglary
- Motor vehicle theft
- Theft

For each of these victimization types, estimates were produced for the following subpopulations:

Personal Victimization Subpopulations
- Sex
- Race
- Age Category
- Region
- Rural/Urban
- MSA Status

Property Victimization Subpopulations
- Household Income
- Region
- Rural/Urban
- MSA Status

To study the relationships among these variance estimates, the percent relative standard error (RSE) was used. The percent RSE is the square root of the variance of an estimate divided by the estimate, and is expressed as a percentage $\left(100 \times \sqrt{Var(Y)}/Y\right)$. The percent RSE removes the scale of the estimate and allows comparisons to be made across multiple types of estimates with different scales (e.g., totals versus rates).

As previously noted, in 2012, BJS shifted from excluding series reported victimizations to including series reported victimizations in NCVS analyses and products. In July 2013, the U.S. Census Bureau released new GVFs for estimates in which series reported victimizations were *included*, whereas previously released GVFs were for estimates in which series reported victimizations were *excluded*. For this reason, separate consideration is given to estimates from 2008 and later versus estimates created prior to 2008.

4.1. 2008-2011 Single Year Estimates

Exhibit 3 presents three figures summarizing the results for crime victimization rates for single year estimates from 2008 through 2011. Series reported victimizations are included in the estimates and the GVFs. The figures display the relationship between the three variance estimation methods—TSL vs. GVF, BRR vs. GVF, and TSL vs. BRR—by plotting percent RSE from one method along the horizontal (x) axis and the alternative method along the vertical (y) axis. If two methods produce consistent results then the bulk of the RSE comparisons would fall along the 45^0 line of equality between the two methods with some estimates varying slightly above or below the line. Figures were also produced for crime victimization totals, but they were almost identical to the victimization rate figures and are therefore, not presented herein.

The first item of note is that both the TSL and the BRR methods match the GVF method well. The RSEs in both ***Figures 1*** and ***2*** are centered on the 45^0 line with no major discrepancies apparent except for a few outlying points. When the RSEs are less than 30%, they are tightly clustered around the 45^0 line, while a wider spread is found for the estimates with RSEs greater than 30%. An estimate with a large RSE is not reliably estimated and will have a wide confidence interval no matter which variance estimation method is used. ***Figures 1*** and ***2*** provide confidence that the TSL and BRR methods applied to the public use data files are matching the methods used by the Census Bureau when producing the GVFs.

A second item of note is that the TSL and BRR methods yield almost exactly the same results as shown in ***Figure 3***. All plotted values are extremely close to the 45^0 line. In addition, the relationship between TSL and BRR was explored for pooled-year estimates and for comparison tests between years. All of these situations also showed that TSL and BRR variance estimates and tests of differences were almost exactly the same for the NCVS public use data. In addition, as described in ***Section 2.2.3***, the BRR method requires a much more complex data set up to account for the phase-in/phase-out of PSUs across the Year Groups than the TSL method. Furthermore, although several analysis packages support both TSL and BRR methods, one of the most widely used by NCVS researchers is SPSS, which does not support BRR variance estimation. For these reasons, BRR direct variance estimation was not examined further, and the remainder of this paper will focus on TSL direct variance estimation.

Exhibit 3. Percent RSEs for Selected Crime Victimization Rates for Single Years from 2008 through 2011

Figure 1. TSL vs. GVF

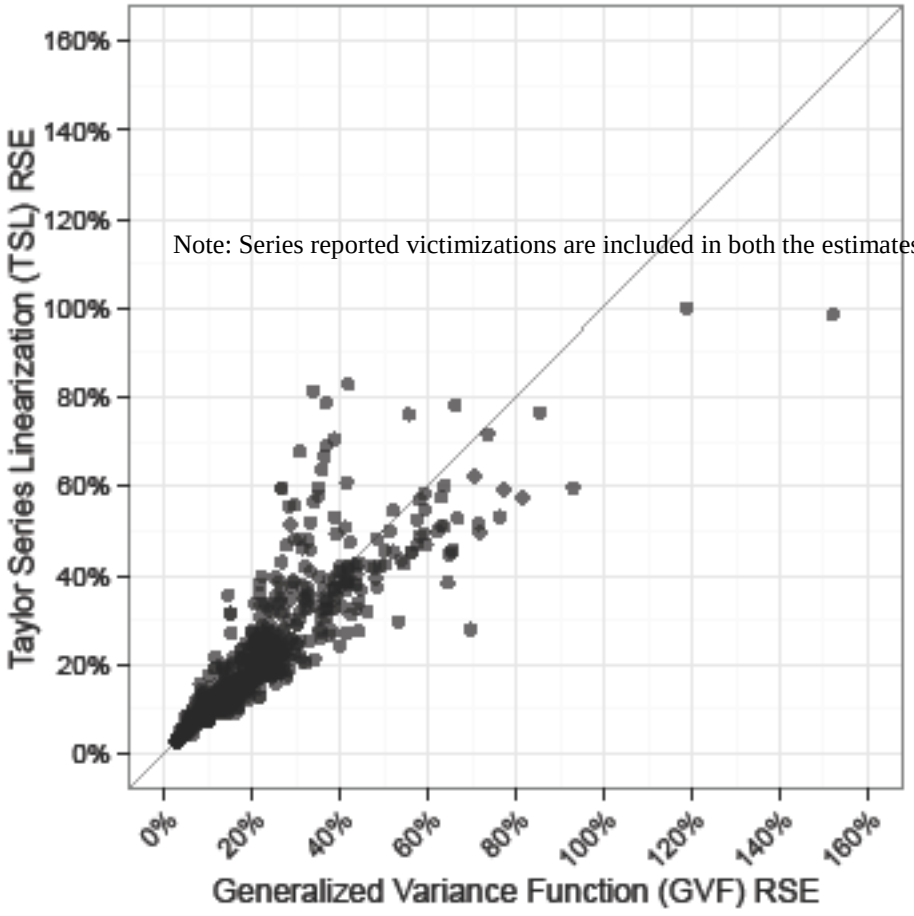


Figure 2. BRR vs. GVF

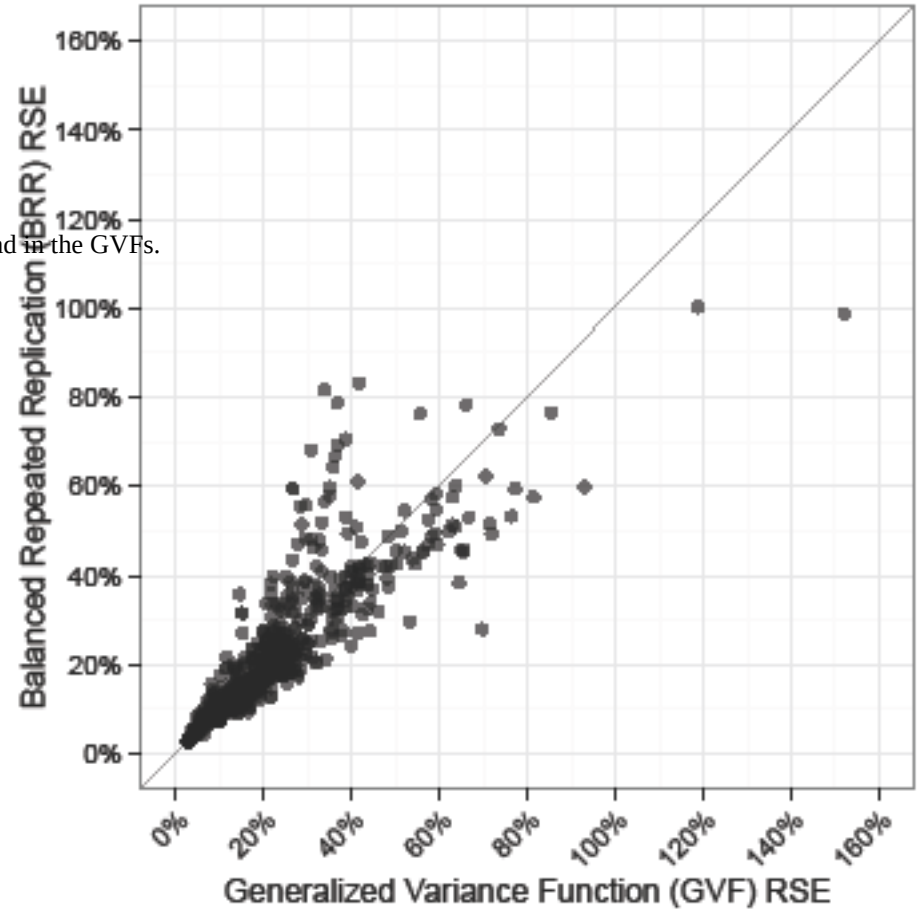
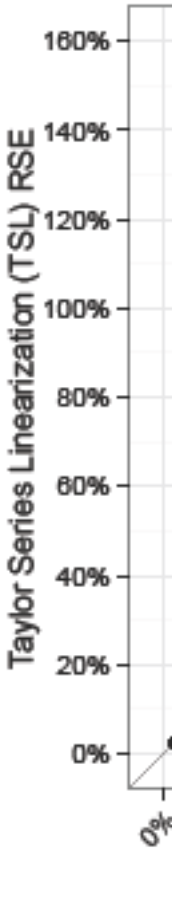


Figure 3. TSL vs. BRR



Note: Series reported victimizations are included in both the estimates and in the GVFs.

4.2. Pre-2008 Single Year Estimates

As stated earlier, for years prior to 2008, the U.S. Census Bureau has not prepared GFVs for estimates in which series victimization reports are included, but GVFs are available for estimates in which series victimizations are excluded. Thus, this section will give special attention to the years prior to 2008 and to the impact that either including or excluding series reported victimizations has on variance estimation.

To explore this situation, single year estimates were prepared for the years 2004 through 2006 both including and excluding series-reported victimizations. Direct TSL variances were calculated for all of these estimates. The GFVs developed excluding series-reported victimizations were applied to all of the estimates, including the estimates in which series victimizations were included. The results are summarized in ***Exhibit 4*** in which the percent RSEs from the TSL method are compared to the percent RSEs from the GVFs.

As demonstrated in a similar analysis presented in ***Section 4.1***, the results for crime victimization totals were almost exactly the same as for rates and have been excluded from this paper. Specifically, when the estimates include series reported victimizations, as shown in ***Figure 1***, the majority of the plotted values are above the 45^0 line of equality, which means that most of the TSL percent RSEs are greater than the GVF percent RSEs. This is likely due to the fact that the GVFs for these years were developed excluding series reported victimizations and the GVF RSEs are too small since they do not account for the added variability that arises from including series-reported victimizations. Additional evidence for this inference is shown in ***Figure 2*** in which the estimates exclude series reported victimizations. In this situation, the TSL and the GVF methods closely align as shown by the tight clustering of the plotted RSEs around the 45^0 line of equality. Although we do not recommend using the GVFs for years prior to 2008 for estimates that include series victimizations, the GVFs for estimates excluding series victimization prior to 2008 appear to be appropriate.

Exhibit 4. Percent RSEs for Selected Crime Victimization Rates for Single Years from 2004 through 2006

Figure 1. Including Series Reported Victimizations

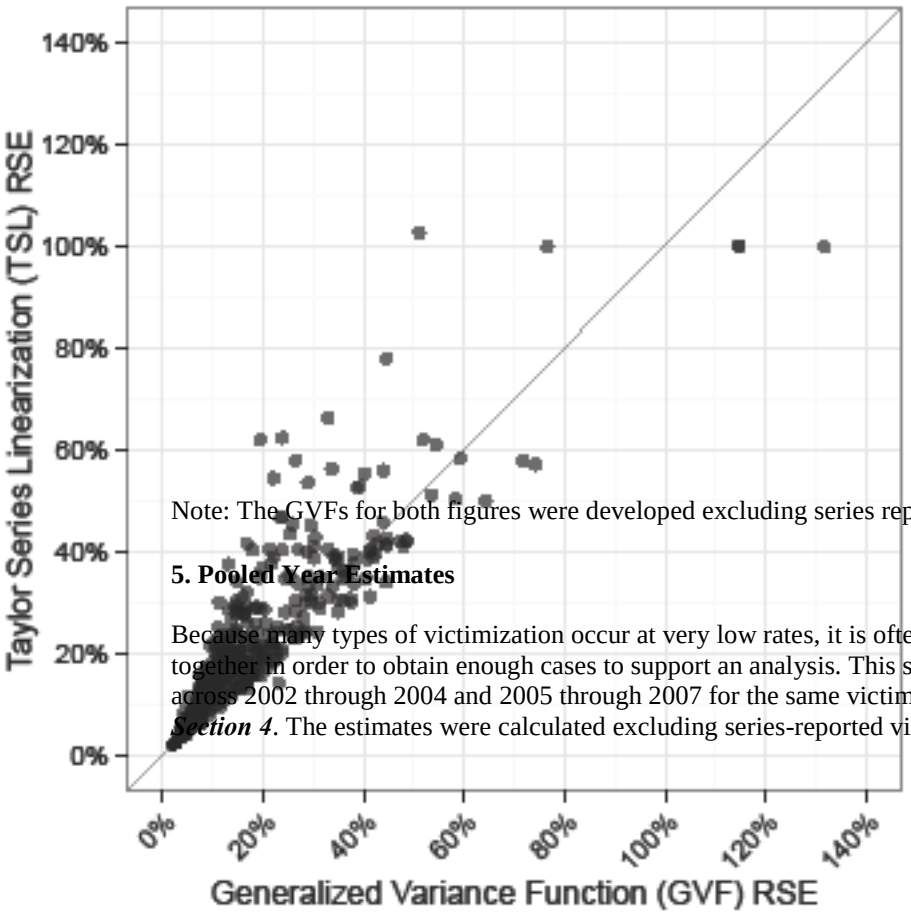
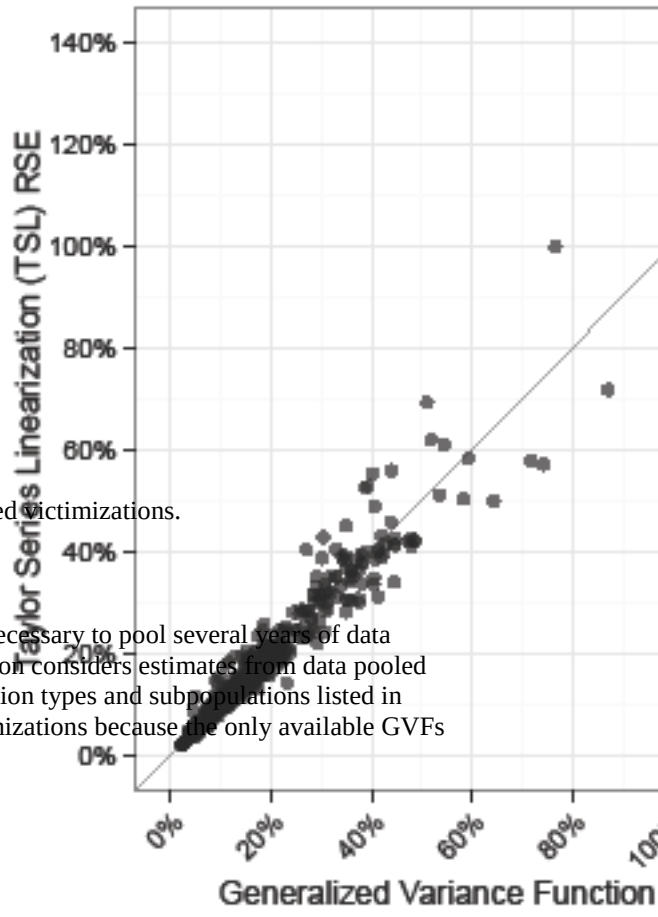


Figure 2. Excluding Series Reported Victimizations



Note: The GVFs for both figures were developed excluding series reported victimizations.

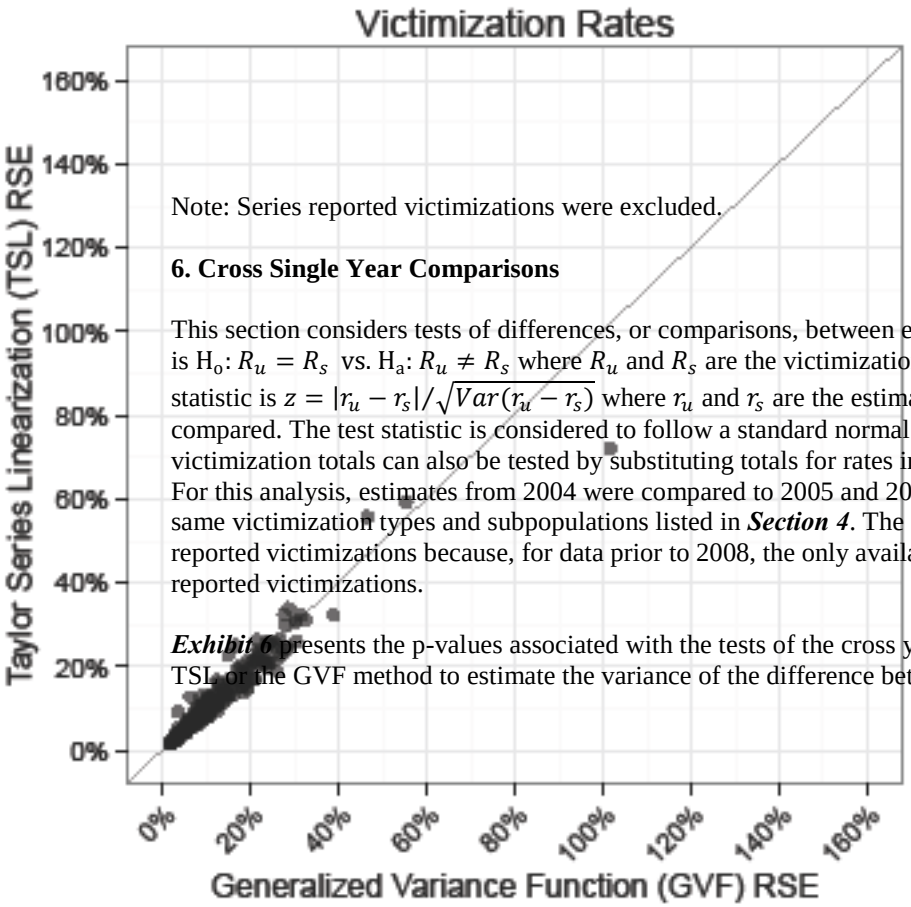
5. Pooled Year Estimates

Because many types of victimization occur at very low rates, it is often necessary to pool several years of data together in order to obtain enough cases to support an analysis. This section considers estimates from data pooled across 2002 through 2004 and 2005 through 2007 for the same victimization types and subpopulations listed in ***Section 4***. The estimates were calculated excluding series-reported victimizations because the only available GVFs

for the years prior to 2008 were created excluding series victimizations. Comparable variance estimates were thus available from both GVF and TSL variance estimation methods for the years under consideration. GFVs created by the Census Bureau that include series reported victimizations are only available for 2008 through 2011, but this four-year window is not long enough to generate non-overlapping three-year time periods for pooled estimates. For this reason, the earlier data from 2002–2004 and 2005–2007 have been used. The TSL direct variance method can be used when either including or excluding series reported victimizations.

Exhibit 5 presents a comparison of the TSL percent RSEs and the GVF percent RSEs for pooled estimates from 2002–2004 and 2005–2007, similar to what was presented in ***Exhibit 3- Figure 1***. The results for crime victimization totals were nearly identical to the rates and thus are not included. The GVF and TSL variance methods correspond very closely for pooled year estimates as demonstrated by the plotted values, which are clustered tightly around the 45^0 line of equality for the two methods. This reinforces the earlier conclusion that the TSL direct variance estimation method has been properly specified for use with the NCSV public use data. In addition, it is expected that pooled year estimates including series reported victimizations will be appropriately addressed by both the TSL and GVF methods as the data become available for the years 2008 and beyond.

Exhibit 5. Percent RSEs for Selected Crime Victimization Rates for Pooled Year Estimates from 2002–2004 and 2005–2007



Note: Series reported victimizations were excluded.

6. Cross Single Year Comparisons

This section considers tests of differences, or comparisons, between estimates from two years. The hypothesis tested is H_o: $R_u = R_s$ vs. H_a: $R_u \neq R_s$ where R_u and R_s are the victimization rates for two different years, u and s. The test statistic is $z = |r_u - r_s| / \sqrt{Var(r_u - r_s)}$ where r_u and r_s are the estimated values of the two victimization rates being compared. The test statistic is considered to follow a standard normal distribution. Likewise, comparisons of victimization totals can also be tested by substituting totals for rates in the preceding hypothesis and test statistic. For this analysis, estimates from 2004 were compared to 2005 and 2005 estimates were compared to 2006 for the same victimization types and subpopulations listed in ***Section 4***. The estimates were computed excluding series reported victimizations because, for data prior to 2008, the only available GVFs were created excluding series reported victimizations.

Exhibit 6 presents the p-values associated with the tests of the cross year comparisons computed using either the TSL or the GVF method to estimate the variance of the difference between two years. Similar to previous exhibits,

the TSL and GVF p-values are compared by plotting the GVF p-values along the horizontal (x) axis and the TSL p-values along the vertical (y) axis. For both victimization rates and totals, the p-values are well aligned along the 45^0 line of equality, which shows that the two methods yield similar results. For victimization totals, there are a few discordant points where the TSL method yields somewhat higher p-values than the GVF method but this does not seem to indicate any systematic discrepancies between the two methods. It is also expected that similar results would have been obtained if series-reported victimizations could have been included with GVFs created including series reported data.

Exhibit 6. P-values for Comparisons between Single Year Victimization Estimates

Note: Comparisons between 2004 estimates vs. 2005 and 2005 estimates vs. 2006. Series reported victimizations were excluded.

7. Cross Pooled Year Comparisons

As was noted in ***Section 5***, it is often necessary to pool several years of data together in order to obtain enough cases to support an analysis. This section extends the discussion in ***Section 5*** to the test of differences, or comparisons, between estimates from two different pooling of years. The same hypothesis and test statistic from ***Section 6*** are considered here, but estimates pooling data from 2002 through 2004 are compared with estimates pooling data from 2005 through 2007 using the same victimization types and subpopulations listed in ***Section 4***. The estimates were computed excluding series-reported victimizations.

Exhibit 7 presents the p-values associate with tests of the cross pooled year comparisons using either the TSL or the GVF variance estimation methods in the same way as was done in ***Exhibit 6***. Again, the TSL and the GVF methods yield similar results for both victimization rates and totals with the p-values well aligned along the 45^0 line of equality. Victimization totals include a few points where the TSL method yields somewhat higher p-values than the GVF method, but a systematic difference between the two methods is not apparent. As before, if GVFs were created including series reported data, it is expected that similar results would have resulted for such data.

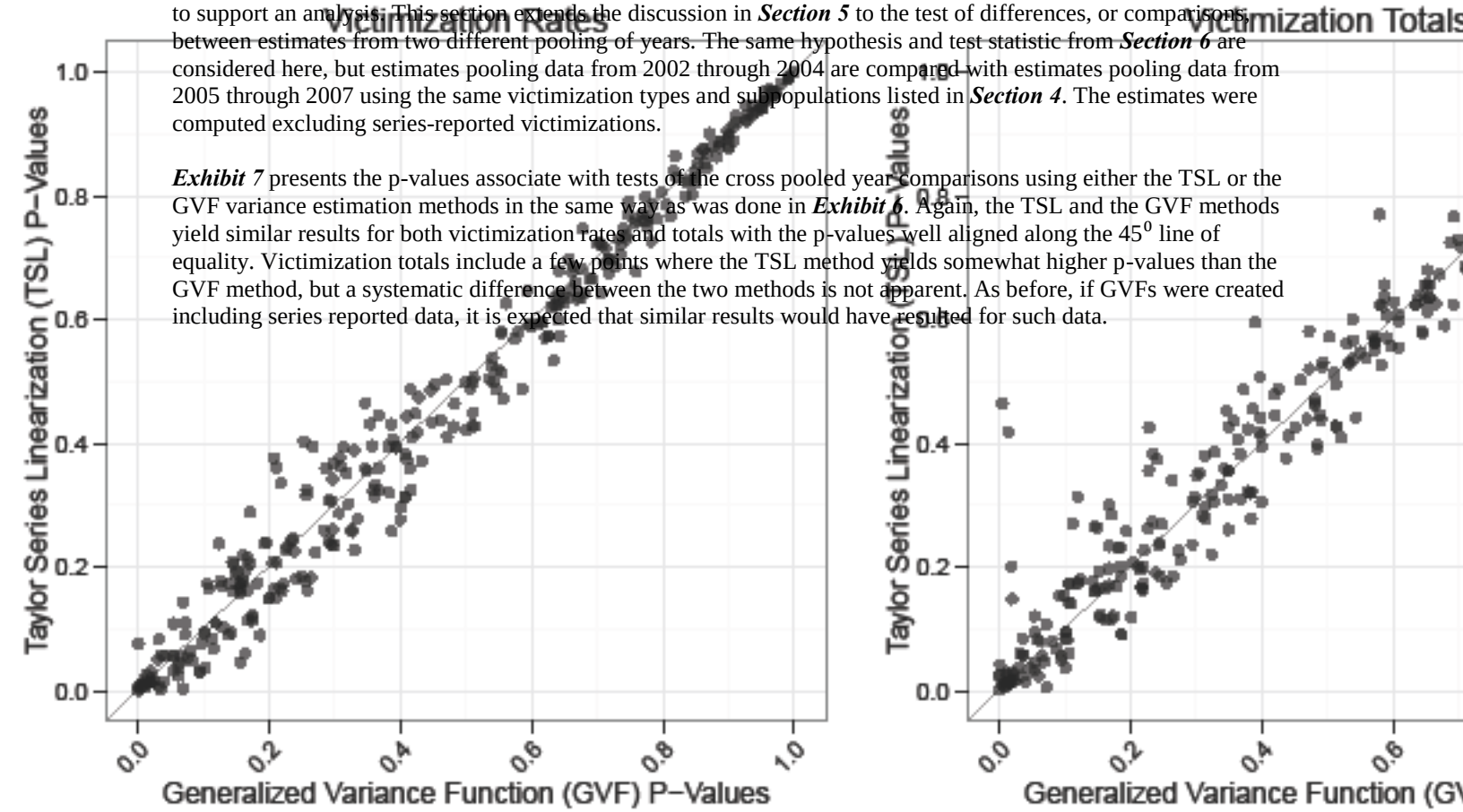


Exhibit 7. P-values for Comparisons between Pooled Year Victimization Estimates

Note: Comparisons between pooled 2002–2004 estimates vs. pooled 2005–2007 estimates. Series reported victimizations are excluded.

8. Discussion

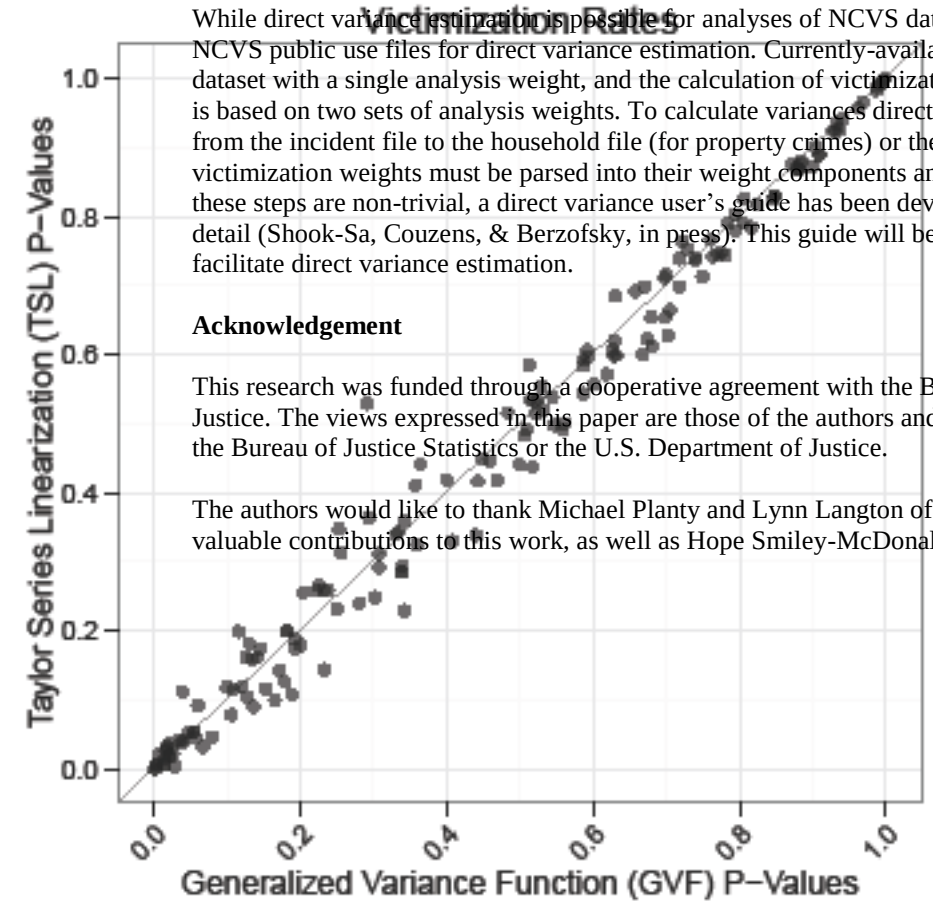
This evaluation found that direct variance estimation techniques can be utilized for the NCVS based on publicly-available data. Comparable results were found between GVFs and direct variance estimates (TSL and BRR), given that the appropriate GVF parameters were used based on the inclusion or exclusion of series victimizations. TSL and BRR produced nearly identical results for single year estimates. Because, for BRR, it is more difficult to prepare analysis datasets and replicate weights, and BRR is not available in the most commonly used software package for NCVS analysts (SPSS), TSL was selected as the most appropriate direct variance estimation method for the NCVS data. GVF and TSL results were comparable for single and pooled year estimates as well as single and pooled cross-year comparisons.

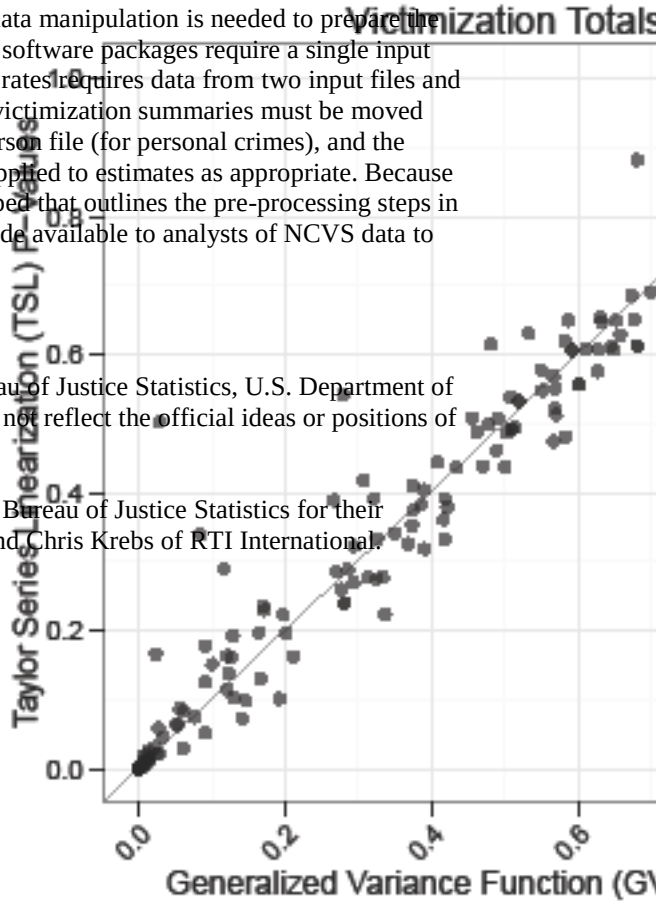
While direct variance estimation is possible for analyses of NCVS data, data manipulation is needed to prepare the NCVS public use files for direct variance estimation. Currently-available software packages require a single input dataset with a single analysis weight, and the calculation of victimization rates requires data from two input files and is based on two sets of analysis weights. To calculate variances directly, victimization summaries must be moved from the incident file to the household file (for property crimes) or the person file (for personal crimes), and the victimization weights must be parsed into their weight components and applied to estimates as appropriate. Because these steps are non-trivial, a direct variance user's guide has been developed that outlines the pre-processing steps in detail (Shook-Sa, Couzens, & Berzofsky, in press). This guide will be made available to analysts of NCVS data to facilitate direct variance estimation.

Acknowledgement

This research was funded through a cooperative agreement with the Bureau of Justice Statistics, U.S. Department of Justice. The views expressed in this paper are those of the authors and do not reflect the official ideas or positions of the Bureau of Justice Statistics or the U.S. Department of Justice.

The authors would like to thank Michael Planty and Lynn Langton of the Bureau of Justice Statistics for their valuable contributions to this work, as well as Hope Smiley-McDonald and Chris Krebs of RTI International.





References

Ash, S. (2010). *Using successive difference replication for estimating variances*. Working paper supplied by US Census Bureau. May be accessed at: http://www.amstat.org/sections/srms/proceedings/y2011/Files/302108_67867.pdf

Cochran, W. G. (1977). *Sampling techniques*. New York: John Wiley & Sons.

Fay, R.E. & Train, G.F. (1995). *Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties*. Joint Statistical Meetings, Proceedings of the Section on Government Statistics, 154-159.

Lauritsen, J.L., Owens, J.G., Planty, M., Rand, M.R., & Truman, J.L. (2012). *Methods for counting high-frequency repeat victimizations in the National Crime Victimization Survey*. Bureau of Justice Statistics Technical Series Report. Bureau of Justice Statistics, Washington, D.C. Available at: http://www.bjs.gov/content/pub/pdf/mchfrv.pdf

Lumley, T. (2008), Balanced repeated replication (BRR). In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods*. Newbury Park, CA: Sage.

Särndal, CE, Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*. Springer-Verlag: New York.

Shook-Sa, B., Couzens, G.L, & Berzofsky, M. (in press). *National Crime Victimization Survey (NCVS) Direct Variance User's Guide*. Prepared for the Bureau of Justice Statistics, Washington, DC.

Williams, R. L. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics, 56*, 218–219.

Williams, R. L. (2008). Taylor series linearization. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods*. Newbury Park, CA: Sage.

Wolter, K.M. (1985), *Introduction to variance estimation*. Spring-Verlag: New York.