The Accuracy of a National Generalized Variance Function for Subnational Estimation

Philip Lee[1], Bonnie Shook-Sa[1], Marcus Berzofsky[1], Lynn Langton[2], Michael Planty[2]
[1]RTI International, 3040 Cornwallis Rd, Research Triangle Park, NC 27709
[2]Bureau of Justice Statistics, 810 Seventh Street, NW, Washington, DC 20531

Abstract

Generalized Variance Functions (GVFs) approximate the variance of an estimate as a function of readily available information about that estimate. They can be used to calculate variance estimates for surveys with complex sample designs, and because they do not require users to have knowledge of the complex design they are often easier to use than direct variance estimation techniques such as Taylor Series Linearization (TSL) for basic analyses. However, the validity of GVFs estimates is only known when they are applied to estimate types that were used to build the GVF equations.

This paper explores the accuracy of a national GVF when applied to subnational estimates using data from the National Crime Victimization Survey (NCVS). The NCVS, sponsored by the Bureau of Justice Statistics (BJS) and conducted by the U.S. Census Bureau, is a multi-mode, rotating panel design survey of households that produces nationally-representative criminal victimization estimates for major types of crimes in the United States. For the NCVS, GVFs created by the Census Bureau were designed to produce variance estimates at the national level, but their accuracy at the subnational level has not been evaluated. We assess the accuracy of GVF estimates within subnational areas based on geographic identifiers on the NCVS Public Use Files (i.e. Census region, population size, and urbanicity) by comparing them with TSL estimates. Our analysis found that TSL and GVFs do not provide consistent variance estimates within these subnational areas and thus, the current NCVS GVFs should not be applied below the national level.

Key Words: GVF, Generalized Variance Functions, NCVS, National Crime Victimization Survey, Taylor Series Linearization, TSL

1. Introduction

Generalized Variance Functions (GVFs) are formulae that approximate the variance of an estimate as a function of readily available information about the estimate (Wolter, 1985). GVFs are developed through a modeling process where variances for a variety of key estimate types are estimated using direct variance estimation methods such as Taylor Series Linearization (TSL), Jackknife, Balanced Repeated Replication, and Successive Difference Replication. The variances are modeled as a function of various values such as the estimate, sample or population size, characteristics related to the sample design and respondent characteristics. Separate models are also developed for various types of estimates such as rates, totals, and percentages. These models take into account the complex sample design, but users of the GVF equations only need the point estimate and other model parameters (e.g. population size, estimate type) to approximate the design-

consistent variance estimates. This makes GVFs appealing for data collections with complex designs that seek to simplify analyses for end users of the data.

An example of a GVF for a rate (from the National Crime Victimization Survey, or NCVS) is:

$$V_r(\hat{r}, \widehat{N}; b, c) = b\frac{\hat{r}(1000 - \hat{r})}{\widehat{N}} + c\frac{\hat{r}\left(\sqrt{1000\hat{r}} - \hat{r}\right)}{\sqrt{\widehat{N}}}$$

where $\hat{r}$ is the point estimate (for a rate) calculated from the survey, $\widehat{N}$ is the estimated population size, and *b* and *c* are GVF parameters calculated in the modeling process and provided to end users. To approximate the variance of an estimated rate, an analyst simply plugs the estimated rate, population size, and GVF parameters into the equation above.

In this paper we examined whether using GVFs designed for a super or parent-level estimate type produces accurate variance estimation in a subdomain analysis through a case study with data from the NCVS. We explore whether GVFs developed at the national level produce reliable variance estimates at the subnational level by comparing subnational GVF estimates to direct variance estimates.

2. Case Study Background: The National Crime Victimization Survey

The NCVS, conducted since 1973 and sponsored by the Bureau of Justice Statistics (BJS), is a nationally representative sample of approximately 50,000 households and 75,000 persons interviewed two times per year. The survey provides estimates of the frequency and characteristics of non-fatal crime victimization in the United States.

The survey is designed to produce only national estimates; however, BJS recognizes the importance of subnational estimates and is exploring various approaches for calculating and disseminating estimates at the subnational level. As part of this investigation, BJS intends to begin developing 'generic area' typologies based on various geographic, social, economic, or demographic characteristics. These generic areas will then represent all places that are similar to each other based on the characteristics of interest.

Three subnational geographic identifiers are available on the NCVS Public Use File (PUF): region, population size, and urbanicity (i.e. location of residence). By crossing these three variables, four two- and three-variable generic area types can be formed (Planty 2012). Thus, crime rates could be reported for each of these generic areas to give stakeholders an indication of the victimization trends in areas like theirs (e.g. rural areas with <100,000-249,999 persons in the South). ***Table 1*** defines the variables found in the NCVS's PUF that are used in the formation of the generic areas.

Table 1: Table of Subnational Geographic Variables Found in the NCVS

Generic Area Variable	Generic Area Variable Levels	Generic Area Variable Source(s)
Census Region[3]	1 = Northeast 2 = Midwest 3 = South 4 = West	Census region classification
Population Size[1, 3]	1 = Not in a place 2 = <100,000 – 249,999 3 = 250,000 – 999,999 4 = 1,000,000 +	Census place size code • 1990 Census population for the 1996-2005 NCVS • 2000 Census population for the 2006-2012 NCVS
Urbanicity[3]	1 = Central or Principal city of a MSA/CBSA (Urban) 2 = in MSA/CBSA, but not in the Central or Principal city (Suburban) 3 = not in an MSA/CBSA (Rural)	CBSA/MSA Status • 1993 MSA and central city classifications for the 1996-2005 NCVS • 2003 CBSA[2] and principal city classifications for the 2006-2012 NCVS

3. Study Methods

Within each generic area, victimization rates and totals, and their accompanying variance estimates were calculated using GVFs designed to be used for national estimates. These GVF estimates were compared to variance estimates produced using direct variance estimation from statistical software that takes into account the complex NCVS sample design. GVF estimates were based on the series-adjusted GVF parameters provided by the US Census Bureau and were aggregated for pooled year estimates using Census-provided correlations based on the crime type. Because direct variance estimation takes into account the sample design, subnational sample sizes, and the weights of cases within each subnational area, direct estimates are known to produce valid variance estimates at the subnational level (Shook-Sa et. al, forthcoming). Thus, they can serve as a gold-standard for evaluating GVFs applied to generic areas. In our analysis, we used SUDAAN software to calculate TSL direct variance estimates, as prior research found TSL to be the most straightforward direct variance estimation approach for the NCVS (Williams et. al, 2014).

For each generic area, victimization rates and totals, and their accompanying variance estimates were computed for the twenty-three crime types presented in ***Table 2***, including both overall crime and those reported to the police.

Table 2: Table of the Twenty-Three Crime Types Used in the Analysis

Type of Crime:		
All crime	**Violent crime**	**Serious violent crime**
Rape/sexual assault	**Robbery**	**Assault**
Aggravated assault	**Simple assault**	**Personal theft**
Violent Crime:		
Violent crimes involving a weapon	**Violent crime involving a firearm**	**Violent crimes committed by a stranger**
Violent crimes committed by an intimate	**Violent crimes committed by other relative**	**Violent crimes committed by other known offender**
Violent crimes occurring during the day	**Violent crimes occurring at night**	
Property Crime:		
Property Crime	**Household Burglary**	**Motor Vehicle Theft**
Theft	**Household crimes occurring during the day**	**Household crimes occurring at night**

These crime types provide a range of estimates from personal to property, and from more common crime types to rarer crime types, thus allowing an evaluation of how GVFs perform on a range of estimate types. The following years were included in the analysis:

- 1-year estimate: 2002, 2008, 2012
- 3-year estimates: 2000-2002, 2006-2008, 2010-2012
- 5-year estimates: 1998-2002, 2004-2008, 2008-2012

For each number of pooled years, 8,694 estimates were computed for both rates and totals across the analysis years, crime types, and generic area types.

Because GVFs and direct variance estimates produce the same point estimates (only the estimated variances differ), one measure that can be used to compare the estimated precision is the percent Relative Standard Error (RSE), calculated as the square root of the variance of an estimate divided by the estimate, expressed as a percentage $\left(100 \times \sqrt{Var(Y)}/Y\right)$. This standardized estimate provides a better comparison when estimates of different types (e.g. rates and totals) are compared. Thus, percent RSEs were calculated for each of the GVF and direct estimates described above.

4. Results

This section presents the results of the GVF evaluation. Section 3.1 evaluates the results by the number of pooled years, and Section 3.2 evaluates the results by crime type.

4.1. Comparison of GVF Variance Estimates to Direct Variance Estimation in Generic Areas by Years Pooled

To summarize the comparisons between GVF variance estimates and direct variances, ***Figure 1*** compares the RSEs produced with GVFs and the RSEs produced by direct variance estimation for victimization rates. From ***Figure 1*** it is clear that the GVFs and direct variance estimates do not track well, given the high levels of deviation from the 45 degree line of equality. For the 1-year victimization rates, the majority of points fall above the 45 degree line, indicating that the GVF approach yielded smaller standard errors in comparison to the direct estimation approach. Thus the GVF method tends to underestimate the variances of 1-year estimates while the direct variance approach produces more accurate variance estimates. Therefore, the use of the GVF for 1-year generic estimates will lead to an increase in the Type I error rate whereby more comparisons will be deemed significantly different than should be. The reverse is true for 3- and 5-year victimization rates and totals, where the majority of points fall below the 45 degree line. This indicates that GVF standard errors are larger than direct estimation standard errors, and thus GVFs are overestimating the variances. Therefore, the use of the GVF for 3- and 5-year generic estimates will lead to an increase in the Type II error rate whereby more comparisons will be deemed statistically similar than should be. The same patterns held when concentrating only on reliable estimates, those with GVF and direct variance RSEs less than 30 percent (results not shown). A similar figure was produced for victimization totals but is not included here because the results were similar.

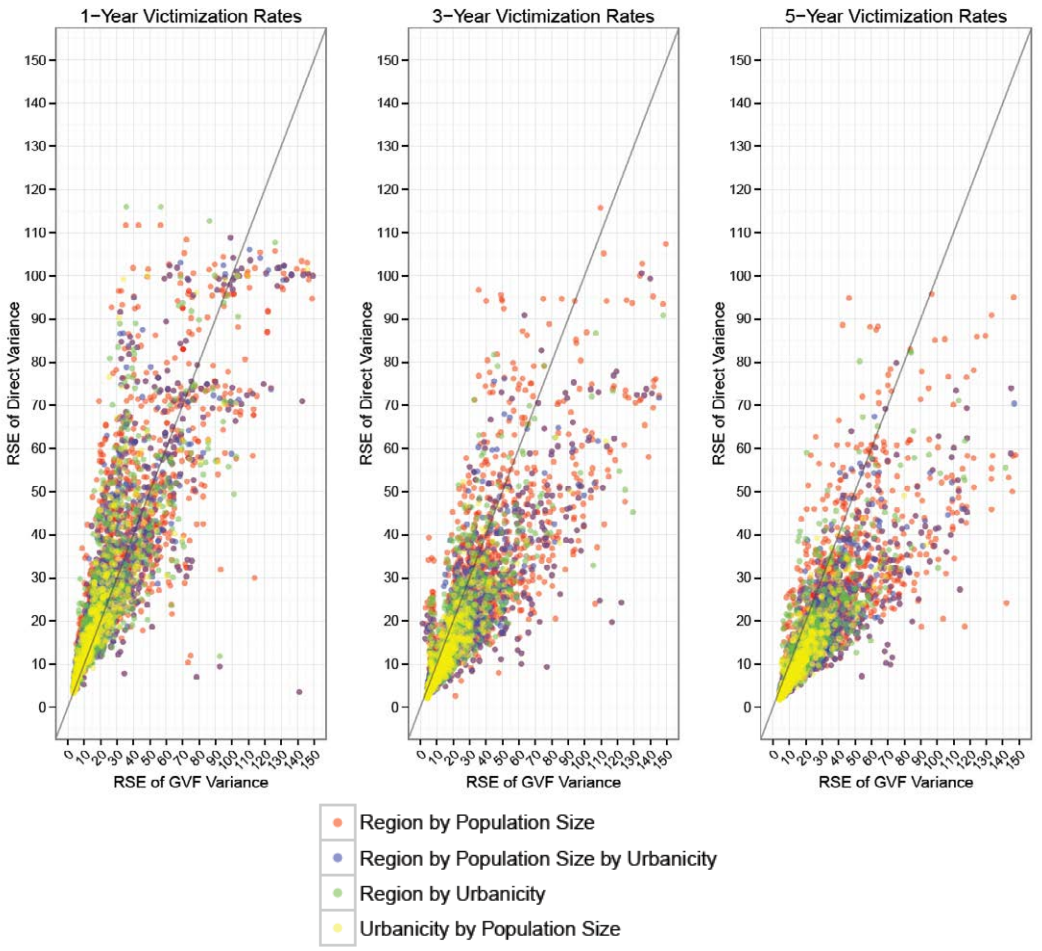

Figure 1: Comparison of GVF RSEs to Direct Variance RSEs: 1-, 3-, and 5- Year Victimization Rates

Table 3 further compares differences across 1-, 3-, and 5-year estimates by displaying the percentage of estimates where the GVF standard error is less than the direct estimation standard error (i.e., the percent of estimates for which the GVF is underestimating the variance). Because GVF estimates are dependent upon the GVF parameters, which vary by year, ***Table 3*** also splits out these estimates by the analysis year. In addition, ***Table 3*** allows for the comparison of estimates across generic area types.

Table 3: Percentage of Estimates where GVF Standard Error is Less Than Direct Variance Standard Error by Year, Year Group and Generic Area Type

		1-year		3-year		5-year	
Generic Area	Year	Totals	Rates	Totals	Rates	Totals	Rates
Region by Population Size by Urbanicity	2002	66.8 %	61.5 %	43.4 %	36.5 %	28.6 %	16.4 %
	2008	54.2	51.3	26.2	15.3	12.7	5.4
	2012	48.4	42.8	25.9	13.8	16.8	5.8
Region by Population Size	2002	61.1	60.3	30	28.3	11.3	8.7
	2008	55.6	55.8	16.6	13.2	3.3	2.6
	2012	48.6	48.8	13.5	12.1	8.2	5.6
Region by Urbanicity	2002	75.2	66.8	43.3	31.9	29.9	17.8
	2008	61.6	57.6	33	15.8	14.7	4
	2012	60.3	53.3	30.6	15.9	21.6	6.3
Urbanicity by Population Size	2002	68.9	66.5	34.8	25.2	21.1	11.8
	2008	62.1	59.9	18	6.8	5.6	1.2
	2012	60.2	57.8	18.6	8.7	13	3.1

If the GVF and direct estimation approaches provided consistent results, then entries in the table would track right around 50 percent. That is, half of the time GVF estimates would track slightly above direct estimates and the other half of the time GVF estimates would track slightly below direct estimates. However, ***Table 3*** demonstrates the same patterns that were present in ***Figure 1*** – 1-year GVF estimates tend to underestimate variances (percentages are greater than 50 percent) while 3- and 5-year estimates tend to overestimate variances (percentages are less than 50 percent).

However, there are clear differences across analysis years, with 2002 GVF variance estimates being smaller than direct estimates more frequently than in 2008 and 2012 for 1-, 3-, and 5-years (i.e., the entries in ***Table 3*** are greater for 2002 than 2008 and 2012). This is in line with something that was noted during the analysis. When calculating GVFs for the 3-year period of 2000 – 2002, there were a total of 84 estimates for rates that had negative GVF variance estimates, all of which were associated with property crimes.

For example, the 2000-2002 estimated rate of motor vehicle theft in Western/rural areas was calculated to be negative based on the GVFs. These negative variance estimates are due to negative GVF parameters in 2000 and 2001 for overall property crime estimates, which produce positive variance estimates at the national level but not within some subnational areas. While this situation is rare (negative variance estimates were computed for only 2.9 percent of victimization rates in the 2000-2002 period), negative variances can be reported when using the GVFs for an estimate type they were not designed to accommodate (e.g. a subnational level estimate). Even though GVF estimates are rarely negative, 2002 GVFs underestimated the true variances more frequently than the other two analysis years.

4.2. Comparison of GVF to Direct Variance Estimation across Crime Type and Generic Area Type

In addition to assessing differences across years and year groups, comparisons were made to assess differences between GVF and direct variance estimates across crime types. ***Table 4*** presents the percentage of estimates for rates and totals where the GVF standard error is less than the direct variance standard error by generic area type, number of years pooled, and type of crime. The table associated with crimes reported to police (not shown) showed similar results.

Table 4: Comparison of GVF and Direct Variance Estimates: Differences by Crime Type for Overall Crimes[1]

Percent of estimates where GVF SE < Direct SE

	Region by Population Size by Urbanicity			Region by Population Size			Region by Urbanicity			Urbanicity by Population Size		
Type of Crime	1-year Rates	3-year Rates	5-year Rates	1-year Rates	3-year Rates	5-year Rates	1-year Rates	3-year Rates	5-year Rates	1-year Rates	3-year Rates	5-year Rates
Overall												
All crime	94.0 %	44.0 %	27.4 %	91.7 %	45.8 %	20.8 %	97.2 %	38.9 %	27.8 %	85.7 %	38.1 %	14.3 %
Violent crime	73.8	39.3	14.3	77.1	31.3	8.3	83.3	30.6	16.7	76.2	14.3	14.3
Serious violent crime	46.4	20.2	7.1	50.0	12.5	0.0	55.6	11.1	5.6	61.9	4.8	0.0
Rape/sexual assault	38.1	23.8	14.3	47.9	25.0	14.6	52.8	33.3	11.1	61.9	38.1	19.0
Robbery	29.8	17.9	6.0	37.5	14.6	4.2	33.3	11.1	11.1	42.9	9.5	4.8
Assault	72.6	36.9	9.5	70.8	29.2	6.3	80.6	33.3	16.7	81.0	14.3	9.5
Aggravated	34.5	22.6	10.7	35.4	25.0	4.2	41.7	19.4	8.3	52.4	23.8	0.0
Simple	70.2	33.3	14.3	75.0	25.0	12.5	80.6	33.3	16.7	85.7	19.0	14.3
Violent crimes involving a firearm	20.2	13.1	3.6	22.9	8.3	4.2	25.0	16.7	2.8	33.3	9.5	4.8
Personal theft	8.3	2.4	0.0	14.6	0.0	0.0	5.6	2.8	0.0	19.0	0.0	0.0
Property crime	94.0	34.5	20.2	93.8	29.2	8.3	91.7	30.6	25.0	90.5	19.0	9.5
Household burglary	86.9	41.7	28.6	89.6	33.3	22.9	94.4	36.1	19.4	95.2	19.0	14.3
Motor vehicle theft	66.7	36.9	16.7	77.1	35.4	10.4	66.7	33.3	8.3	76.2	28.6	9.5
Theft	88.1	32.1	15.5	85.4	29.2	10.4	86.1	25.0	16.7	85.7	14.3	9.5

[1] Note: Not all crime types are included

This table shows that for some crime types the GVFs tend to overestimate the variance, while for others they tend to underestimate the true variance. As previously noted, 1-year GVFs tend to underestimate variances, but this pattern does not hold for all crime types. For example, GVFs tend to overestimate variances for robbery and personal theft. One example of a crime type where the GVFs always overestimate the variance is violent crimes involving a firearm – where the GVFs are too high. While some crime types exhibit clear patterns, the majority of crime types vary in whether they overestimate or underestimate the variance depending on the number of pooled years and the generic area type. ***Table 4*** also support the conclusions made from ***Table 3*** and ***Figure 1*** that the more years that are pooled, the more the GVFs overestimate variances (as is evident by the decreasing percentages as more years are pooled within each generic area).

Based on this analysis, it is clear that the direct estimation and GVF approaches do not provide consistent variance estimates at the generic area level. Because GVFs were only designed to produce national estimates and direct estimation has been validated for use on the NCVS previously, this provides evidence that GVFs are not accurate at the subnational level.

5. Conclusions

GVFs are a useful tool to allow end users of complex samples to calculate design-consistent variance estimates without knowledge of the complex sample design. However, the validity of GVF estimates is only known for estimate types included in the development of the GVF models. This paper explored the accuracy of a national GVF at the subnational level and found that GVFs did not track well with direct variance estimates. The GVFs evaluated tended to underestimate variances of single-year estimates and overestimate variances of pooled year estimates. Furthermore, when GVFs were inappropriately applied at the subnational level, negative variance estimates were sometimes produced.

While this analysis focused on applying national GVFs to the subnational level, similar problems can occur when GVFs are applied to other estimate types not included in the model (e.g. estimates for subpopulations not included in the model, different analytic years). Whenever possible, direct estimation procedures provide the most accurate variance estimates when the design is known and can be specified at the time of estimation. When GVFs are applied, care should be taken to only apply GVFs to the types of estimates included in the GVF model development. Reliability should not be assumed for any other estimate types.

Acknowledgements

We would like to note that the views expressed in this paper are those of the authors only and do not reflect the views or position of BJS or the Department of Justice.

References

Planty, M. (2012). Approaches to estimating subnational victimization estimates. In *Federal Committee on Statistical Methodology Proceedings*, pp. 1–10. Retrieved from http://fcsm.sites.usa.gov/files/2014/05/Planty_2012FCSM_I-B.pdf

Shook-Sa, B., Berzofsky, M.E., Couzens, L., Moore, A.., and Lee, P. (2015). *Assessing the Coverage and Reliability of NCVS Sample in the Largest States, MSAs, and Cities.* Prepared for the Bureau of Justice Statistics, Washington, DC.

Williams, R., Heller, D., Couzens, G. L., Shook-Sa, B., Berzofsky, M., Smiley-McDonald, H., & Krebs, C. (2014). *Evaluation of direct variance estimation, estimate reliability, and confidence intervals for the National Crime Victimization Survey*. Prepared for the Bureau of Justice Statistics, Washington, DC.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.