Generic Areas and Small Area Estimates for Subnational Estimation on the National Crime Victimization Survey

Philip Lee[1], Marcus Berzofsky[1]
[1]RTI International, 3040 Cornwallis Rd, Research Triangle Park, NC 27709

Abstract
The National Crime Victimization Survey has historically produced nationally-representative criminal victimization estimates for major types of crime in the US. However, there is interest in understanding local level crime problems and examining state and local variations in crime rates. To address this demand, BJS is developing a portfolio of both direct and indirect estimation procedures. Two broad methods that can be used to produce subnational estimates from the existing sample when sample is insufficient for reliable estimates: (1) Generic Areas and (2) Small Area Estimates (SAE). Generic areas can be constructed using variables available on the NCVS public use file (region, population size, and urbanicity) allowing users to dynamically produce subnational estimates; however, estimates for specific areas are unknown. SAEs use existing data to model subnational estimates for specific areas. SAEs offer victimization rates for a specific state or MSA but are static estimates. We discuss the advantages and disadvantages of each method. In addition, through graphics, we demonstrate and show examples of the types of estimates produced under each approach.

Key Words: Generic areas, Small Area Estimates, SAE, NCVS, National Crime Victimization Survey

1. Background

The National Crime Victimization Survey (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 has historically produced nationally-representative criminal victimization estimates for major types of non-fatal crime in the United States.

Prior to 2016, the NCVS was designed to only produce nationally representative estimates, but not representative for smaller areas such as states, cities, and Core-Based Statistical Areas (CBSAs). Data to analyze the NCVS is provided in two formats, a Public Use File (PUF) and a Census-restricted dataset. The difference between the two datasets is that the PUF does not contain any geographic information below Census region, such as states and counties. This is intentional to protect the confidentiality of the respondents. Access to Research Data Centers (RDCs) and Census Headquarters are required to access the geographic data for the NCVS.

2. Motivation and Research Questions

BJS recognizes that there is interest in understanding local level crime problems and examining state and local variations in crime rates. Currently, the main source for local crime rates is the FBI's Uniform Crime Reporting (UCR). The UCR provides estimates based on law enforcement agencies and provide summary reporting of the victimization; however, the rates provided are based on crimes *reported* to the police. There is no source that allows analysts to understand reported and non-reported crime at the local level other than the restricted-use NCVS. Thus, BJS is exploring various approaches to improve user's ability to analyze data at the subnational level. As part of this investigation, BJS intends to develop 'generic area' typologies based on three subnational geographic identifiers that are available on the NCVS's PUF: region, population size, and urbanicity (i.e. location of residence). In addition, BJS has started developing small area estimates (SAE) which uses statistical methods to develop model based subnational estimates for specific areas such as states and various large places.

BJS have been working on subnational victimization estimates for many years and the two current methods for subnational estimation are generic areas and SAE, both of which are developed independently. No research has been done to compare the estimates produced from both methods at a subnational level. For our research, we asked the following questions:

1. Do generic areas and small area estimates produced similar estimates for violent and property crimes in comparable areas?
2. If estimates are different, are they systematically different (e.g., SAE are usually higher or lower than generic area)?

3. Review of Approaches for Producing Subnational Estimates

This section presents information of how the two methods work and provide advantages and disadvantages of using each method.

3.1 Generic Areas

The idea of the generic area approach is to create a typology of 'like places'. Many PUF must remove geo-identifiers for disclosure reasons, but generic areas can be constructed using variables available on the PUF (in this case, Census region, population size, and urbanicity) allowing users to dynamically produce subnational estimates for say, urban areas with a population of 1 million or more in the Midwest; however, estimates for specific areas are unknown (Planty 2012).

Using publicly available information on the PUF version of the NCVS. Create "generic" areas through the combination of region, population size, and urbanicity. Four types of generic areas could be produced: region by population size, region by urbanicity, population size by urbanicity, and region by population size by urbanicity. An example of how generic area compares to actual subnational area would be New York City, NY could be generalized as an urban area in the Northeast region, where there are over 1,000,000 people.

Table 1 below lists the advantages and disadvantages of using generic areas in the NCVS.

Table 1: Advantages and Disadvantages of using Generic Areas in the NCVS

Advantages	Disadvantages
Do not need auxiliary source. Generic areas can be produced using the variables currently in the NCVS, i.e. Census region, population size, and urbanicity, that's PUF variables V2127B, V2126B, and V2129 respectively.	**One year estimates less reliable**. Smaller sample sizes mean less precision for the 1 year estimates compared to 3 or 5 year aggregate estimates.
No need to reweight data. The NCVS weights can be used for analysis allowing the estimates to be easily reproduced by other researchers. An analysis of alternative weights adjusted to match Census control totals within each generic area did not have substantive differences in key crime estimates in our investigation (Shook-Sa, Lee, & Berzofsky 2015).	**Rare crime types difficult to measure**. Some of the rarest crime types (e.g., rape and sexual assault, personal theft, violent crimes committed by other relatives) do not have adequate precision for the majority of generic areas and would require further pooling more than 5 years of data or further collapsing of generic area levels to produce estimates with reasonable quality.
Do not require large number of years of data. In general, 3-year pooled estimates have reasonable precision for most types of crime for the two-variable generic area types (Census region by population size, Census region × urbanicity, and population size by urbanicity). Pooling 5 years of data is required to obtain reasonable precision for many crime types for the Census region by population size by urbanicity generic areas.	**NCVS's Generalized Variance Functions cannot be used**. Generalized Variance Functions (GVF) designed for the NCVS at the national level cannot be used to calculate the variance for generic areas. Direct variance estimations using computerized software is recommended unless new GVF are produced specially for generic areas.

3.2 Small Area Estimates

SAEs utilize restricted-use, geo-identified data housed at Research Data Centers (RDCs) and Census Headquarters, along with auxiliary data to model subnational estimates for specific areas. SAEs offer victimization rates for specific states or MSAs but are static estimates that do not allow for regression or other complex analyses.

The SAE statistical models rely on information from the NCVS and auxiliary data sources, such as the FBI's UCR, to model the crime rate patterns and trends across states to produce estimates for all 50 states and select local areas. The NCVS's SAE model requires two sources of information.

1) The nationally representative NCVS dataset with geo-identifiers to identify respondents in subnational areas.
2) An auxiliary source, the UCR, that provides information related to the outcome of interest at the state or local level.

In the following page, **Table 2** below lists advantages and disadvantages of using SAE in the NCVS.

Table 2: Advantages and Disadvantages of using SAE in the NCVS

Advantages	Disadvantages
Auxiliary data helps produce stable estimates in small areas (e.g. counties). Using UCR as an auxiliary data, in some small areas where the NCVS had small sample or no sample at all, credible estimates could still be produced. From research, violent crime estimates produced via SAE are good (Fay and Diallo 2015).	**Areas with little actual NCVS data had large standard errors.** Generally, in smaller states that account for a small portion of the NCVS sample or none of the sample, the estimates largely depends on the UCR data and have large standard errors that will make tests statically insignificant.
Produce estimates and examine trends for all 50 states. Using UCR as an auxiliary data and the NCVS, estimates for all 50 states could be produced and compare to previous years for trend analysis.	**Potential measurement error between NCVS and auxiliary source.** Differences in how the NCVS and UCR measure crime exist. Some key examples include: NCVS does not capture murder, but includes rape, sexual assault, robbery, aggravated assault and simple assault. The UCR does not capture simple assault crimes as a violent crime, but captures murder and nonnegligent manslaughter, forcible rape, robbery and aggravated assault. Location of reported victimization is different. NCVS is based on victim's location of residence whereas UCR is where the crime occurred. For property crimes rates, the NCVS denominator is U.S. households, whereas the UCR is the U.S. population.
3-year estimates possible. Generally, good to use for 3 year estimates for many crime types.	**Difficult to reproduce.** The models are not transparent and computational intensive, and need access to Census HQ or the RDC to work with the confidential Census files.

4. Methods for Comparing Generic Areas to Small Area Estimates

This section presents the method used to compare the generic area estimates (GAE) to SAE. The 3-year SAE came from the report "Developmental Estimates of Subnational Crime Rates Based on the National Crime Victimization Survey" by Fay and Diallo (2015). In their report, SAE were produced for the 50 states and D.C., selected large counties, and CBSAs. The GAE came from pooling 5 years worth of NCVS's PUF to obtain reasonable precision for the three-way cross of Census region, population size, and urbanicity generic areas as recommended from the report "Assessing the Coverage and Reliability of Subnational Geographic Identifiers in the NCVS Public-Use File" by Shook-Sa, Lee, and Berzofsky (2015)[1].

To compare estimates for both subnational estimation methods, a comparable reference year and specific set of areas must be identified. For this analysis, the reference year 2010 was chosen as Census's population totals are available to the public and estimates from the SAE report (Fay and Diallo, 2015) are available. Thus, for SAE the reference period used was 2009 through 2011 from the SAE report and GAE used NCVS's PUF from 2008 through 2012 to create the 3 and 5 year rolling averages respectively that contains 2010 as the reference period. Counties were chosen as county-level information is readily available from U.S. Census Bureau's website regarding population totals and region for the reference period 2010; counties were chosen since they are the smallest subnational area used in the SAE report and more rational to generalize to one urbanicity type when compared to larger areas such as states and CBSAs. The restricted-use NCVS datasets were used to determine which of the 65 counties, from the SAE report, were primarily urban based on the NCVS's household dataset. Based on our review, 29 of the 65 counties are identified as majority urban and, for this analysis, were assumed to be completely urban for the generic area classification[2].

5. Results

To compare the GAE and SAE estimates, we produced a plot of the total violent victimization rates per 1,000 persons and the total property victimization rates per 1,000 households in 2010 for the three-way cross GAE of Census region, population size, and urbanicity and the SAE for the 29 largest urban counties. For each plot, the x-axis displays the GAE rates for 5-year average for years 2008-2012. The y-axis displays the SAE rates for 3-year average for years 2009-2011. The diagonal line in the figures is the identity line y=x (i.e., the rate by which the GAE and SAE estimates are equal). Since the GAE rates are on the x-axis, a vertical line is drawn to indicate urban counties with the same GAE rates.

[1] Research conducted after this report was published indicated using a rolling 3-year average for GAEs produces reliable estimates for total violent and total property crimes.
[2] While most counties are a mix of urbanicity types, for this analysis, a county was assigned one designation.

5.1 Total Violent Crime

Figure 1 presents the comparison between GAE and SAE for total violent crime victimization. Points that falls below the identity line indicate that GAE are larger than the SAE, while points that fall above the identity line indicate that GAE are lower than SAE. From **Figure 1**, GAE produced estimates that are larger than SAE for these 29 urban counties, since most of the points fall under the identity line. **Table 4** breaks down **Figure 1** by generic areas to show the number of counties where GAE are greater than the SAE and the distribution of the differences in the total violent rates between GAE and SAE. Negative numbers in **Table 4** on the distribution portion of the table indicates the SAE are larger than the GAE, whereas a positive number indicates that GAE are larger than SAE.

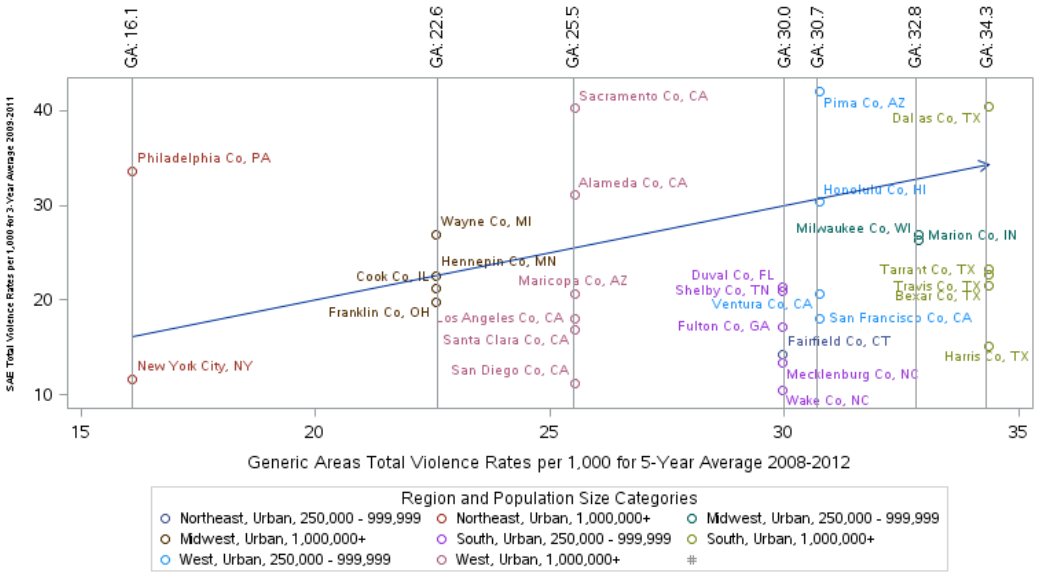

Figure 1: Total Violent Crime Rates per 1,000 Persons in 2010: Generic Areas Estimates (Region by Population Size by Urbanicity) and Small Area Estimates for the 29 Largest Urban Counties

Table 4: Number of Counties Where GAE are Greater Than SAE and Distribution of the Differences in Rates between GAE and SAE for the 29 Largest Urban Counties on Total Violent Rates by Generic Areas

Generic Areas	Number of counties where GAE are higher than SAE	GAE of Total Violent Crimes	Difference between GAE and SAE from Smallest to Largest					
Midwest, Urban, 1,000,000+	2/4	22.6	-4.34	-0.04	1.36	2.76		
Midwest, Urban, 250,000 - 999,999	2/2	32.8	5.95	6.55				
Northeast, Urban, 1,000,000+	1/2	16.1	-17.51	4.49				
Northeast, Urban, 250,000 - 999,999	1/1	30.0	15.67					
South, Urban, 1,000,000+	4/5	34.3	-6.06	11.14	11.64	12.84	19.24	
South, Urban, 250,000 - 999,999	5/5	30.0	8.55	8.95	12.75	16.55	19.45	
West, Urban, 1,000,000+	4/6	25.5	-14.67	-5.57	4.83	7.53	8.63	14.33
West, Urban, 250,000 - 999,999	3/4	30.7	-11.26	0.34	10.14	12.74		

Table 4 shows all generic areas have at least half of the counties' GAE are reported larger than the SAE. Examining **Figure 1** and **Table 4** more thoroughly, some patterns emerge. For example, counties categorized as Midwest, Urban, 1,000,000+ generic areas produce SAE which are relatively close to GAE as the largest absolute difference is 4.34 violent victimization per 1,000 persons or within 20% difference of the GAE. In addition, there are also counties where the GAE are always greater than the SAE. Midwest, Urban, 250,000 - 999,999 generic areas, South, Urban, 250,000 - 999,999 generic areas, and Northeast, Urban, 250,000 - 999,999 generic areas all have counties where the GAE are at least 5.95 violent victimizations per 1,000 persons larger than SAE and have a maximum difference of 19.45 violent victimizations per 1,000 persons.

The remaining generic areas show a mixture of where GAE are less than or greater than the SAE. The Northeast, Urban, 1,000,000+ generic area, which includes Philadelphia county, PA and New York City, NY, reported a SAE of 17.51 violent victimizations per 1,000 persons larger than GAE for Philadelphia, but, on the other hand, reported a SAE 4.49 violent victimizations per 1,000 persons smaller than GAE for New York City.

South, Urban, 1,000,000+ generic area has one county, Dallas county, TX, where the SAE is 6.06 violent victimizations per 1,000 persons larger than GAE. The remaining 4 counties have the GAE at least 11 violent victimizations per 1,000 persons larger than SAE. Tarrant county, TX, Travis county, TX, and Bexar county, TX reported a GAE around 11-13 violent victimizations per 1,000 persons larger than SAE. Within the South, Urban, 1,000,000+ generic area Harris county, TX has the largest absolute difference with a GAE of 19.24 violent victimizations per 1,000 persons larger than SAE.

West, Urban, 1,000,000+ generic area appears to have counties uniformly distributed with Sacramento county, CA reported a SAE that is 14.67 violent victimizations per 1,000 persons larger than the GAE and on the other end San Diego county, CA reported a SAE that is 14.33 violent victimization per 1,000 persons smaller than the GAE. The remaining 4 counties' differences in total violent rates between GAE and SAE appear to be evenly distributed between the above-mentioned differences in total violent victimization.

West, Urban, 250,000 - 999,999 generic area appears to be mixed in terms of SAE with some counties reported higher or lower estimates than GAE.

Pima county, AZ reported a SAE that is 11.26 violent victimizations per 1,000 persons larger than the GAE; Honolulu County, HI's SAE is relatively close to GAE as there is a difference of 0.34 violent victimizations per 1,000 persons between the two estimates; both Ventura county, CA and San Francisco County, CA reported SAE smaller than GAE with a difference of 10.14 violent victimizations per 1,000 persons and 12.74 violent victimizations per 1,000 persons respectively.

5.2 Total Property Crime

Figure 2 compares the total property rates produced by GAE and SAE. Similar to violent crimes, **Figure 2** shows GAE produced estimates that are larger than SAE, since most of the points fall under the identity line. **Table 5** breaks down **Figure 2** by generic areas to show the number of counties where GAE are greater than SAE and the distribution of the differences in the total property crime rates between GAE and SAE. Negative numbers in **Table 5** on the distribution portion of the table indicates that the SAE are larger than the GAE, whereas a positive number indicates that GAE are larger than SAE.

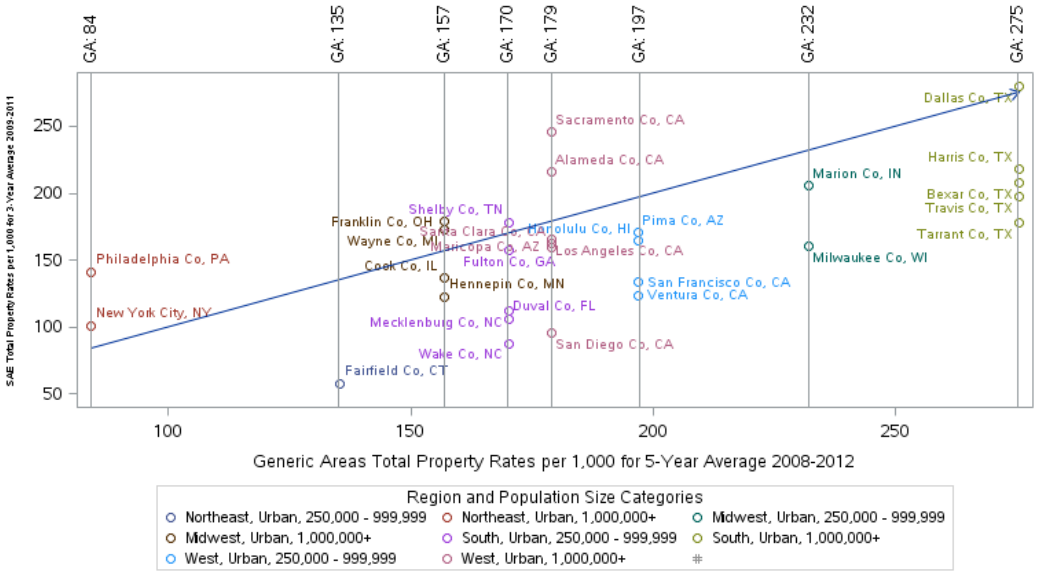

Figure 2: Total Property Crime Rates per 1,000 Households in 2010: Generic Areas Estimates (Region by Population Size by Urbanicity) and Small Area Estimates for the 29 Largest Urban Counties

Table 5: Number of Counties Where GAE are Greater Than SAE and Distribution of the Differences in Rates between GAE and SAE for the 29 Largest Urban Counties on Total Property Rates by Generic Areas

Generic Areas	Number of counties where GAE are higher than SAE	GAE of Total Property Crimes	Difference between GAE and SAE from Smallest to Largest					
Midwest, Urban, 1,000,000+	2/4	157	-22	-16	20	35		
Midwest, Urban, 250,000 - 999,999	2/2	232	26	72				
Northeast, Urban, 1,000,000+	0/2	84	-57	-17				
Northeast, Urban, 250,000 - 999,999	1/1	135	77					
South, Urban, 1,000,000+	4/5	275	-5	57	67	78	97	
South, Urban, 250,000 - 999,999	4/5	170	-8	13	58	64	83	
West, Urban, 1,000,000+	4/6	179	-67	-37	13	17	20	83
West, Urban, 250,000 - 999,999	4/4	197	26	32	63	74		

From **Table 5**, with the exception of the Northeast, Urban, 1,000,000+ generic area, generic areas have at least half of the counties' GAE are reported larger than the SAE. Examining **Figure 2** and **Table 5** more thoroughly, some patterns emerge. For example, like violent victimization, for counties categorized as Midwest, Urban, 1,000,000+ generic area, the SAE produced are relatively close to GAE as the largest absolute difference is 35 property victimizations per 1,000 households or within 23% difference of the GAE. Unlike violent victimization, all counties in the Northeast, Urban, 1,000,000+ generic area reported GAE smaller than the SAE of at least 17 property victimizations per 1,000 households and a maximum difference of 57 property victimizations per 1,000 households.

There are also cases where GAE are always larger than SAE at the county level. Midwest, Urban, 250,000 - 999,999 generic area, Northeast, Urban, 250,000 - 999,999 generic area, and West, Urban, 250,000 - 999,999 generic area all have counties where the GAE are at least 26 property victimizations per 1,000 households larger than SAE and have a maximum difference of 77 property victimizations per 1,000 households.

The remaining generic areas showed a mixture of where GAE are lower or larger than the SAE. Similarly, to violent victimization, the South, Urban, 1,000,000+ generic area has one county, Dallas county, TX, where the SAE reported a larger property victimization rate than the GAE. The remaining 4 counties have the GAE of at least 57 property victimizations per 1,000 households larger than SAE and have a maximum difference of 97 property victimizations per 1,000 households.

The same case also applies to the South, Urban, 250,000 - 999,999 generic area, where one county, Shelby county, TN, where the SAE reported a larger property victimization rate than the GAE. The remaining 4 counties have the GAE at least 13 property victimizations per 1,000 households larger than SAE. Fulton county, GA reported a GAE that is 13 property victimizations per 1,000 households larger than SAE. The remaining counties have reported larger difference with Mecklenburg county, NC, Duval county, FL, and Wake county, NC reported a GAE that is 58 property victimizations per 1,000 households, 64 property victimizations per 1,000 households, and 83 property victimizations per 1,000 households larger than the SAE, respectively.

West, Urban, 1,000,000+ generic area appears to be mixed in terms of SAE with some counties reported higher or lower estimates than GAE. Sacramento county, CA and Alameda county, CA reported a SAE that is 67 property victimizations per 1,000 households and 37 property victimizations per 1,000 households larger than the GAE, respectively. Los Angeles county, CA, Santa Clara county, CA, and Maricopa county, AZ have SAE that are relatively close to the GAE with the GAE of 13 property victimizations per 1,000 households, 17 property victimizations per 1,000 households, and 20 property victimizations per 1,000 households larger than their reported SAE. San Diego county, CA has the largest absolute difference within this generic area with a GAE of 83 property victimizations per 1,000 households larger than the reported SAE.

6. Discussion

From the figures and tables in the previous section, for both total violent and total property crime, the results showed that GAE produced larger estimates compared to the SAE for the 29 urban counties. This section will discuss some possible explanation as to why this is the case.

6.1 Why Generic Areas Produced Larger Total Violent Crime Rates

Although there are various reasons why GAE are larger than SAE for violent crimes, one possible reason is instrument differences in the FBI's UCR for key violent victimization when used as an auxiliary source for the NCVS. As mentioned in the SAE's disadvantage column of **Table 2**, the UCR and NCVS measured violent and property crimes differently, thus the potential for measurement error is possible. One of the key components for total violent crime in the NCVS is the utilization of simple assault crimes. Unfortunately, the UCR does not capture simple assaults, but does captures fatal crimes such as murder. The NCVS however produced statistics on non-fatal crimes, so adjustments were made to the UCR to account for these differences for the SAE. Another difference between the NCVS and UCR is that the UCR only capture reported crime to police, whereas one of the main benefits of the NCVS is the ability to capture crimes that are both reported and not reported to police. Since the UCR does not capture unreported crimes, the potential of under-estimating violent victimization is possible when creating statistics for the total violent crime rates in the NCVS which includes both reported and unreported violent victimization.

6.2 Why Generic Areas Produced Larger Total Property Crime Rates

One key difference between the NCVS and UCR in the calculation of property crime rates is the population of interest. For property crime rates, the NCVS uses U.S. households and the UCR uses U.S. population as the denominator. The potential for measurement error is possible since the population of interests are different for the two instruments if the differences are not accounted for. Another reason why the GAE are larger than SAE for property crime rates for the 29 selected counties used in this analysis have to do with our assumption that if the county is majority urban then assume the corresponding generic area where the urbanicity is completely urban. When producing the SAE for counties, both urban and non-urban areas were used in calculating the rates for property crimes. These 29 selected counties were composed of portions of urban and suburban areas in the makeup of the county. In our analysis, using 5 years worth of NCVS's PUF data from 2008 through 2012, urban areas had larger property crime rates than suburban areas. By assuming that these 29 counties are completely urban, our GAE are over-estimating the property crime victimization. In general, this could also be said of violent victimizations as well regarding why GAE produced larger total violent and total property crime rates than SAE.

7. Limitation and Next Steps

During our analysis, we realized we couldn't compare the remaining 36 counties that are primary suburban in SAE report. The NCVS's suburban status has a maximum population size of 250,000. The issue is that the smallest counties that are considered suburban in the SAE's report have at least 800,000 people in 2010 and is aggravated between urban and non-urban portion of the county. Because there are no suburban counties that have a population size of less than 250,000 in the SAE report, an analysis using a three-way cross generic area for the suburban areas was not done.

Further research is required to create estimates for counties to account for urban and non-urban areas. One method would be to use the two-way cross generic areas of Census region and urbanicity without considering population size to create estimates for each combination. Using the restricted NCVS datasets or other public use sources, we can estimate the proportion for which a county is rural, suburban and urban for each of the 65 counties. With this information, we can create a weighted average using the two-way cross generic areas of Census region and urbanicity and the portion of urbanicity of each county. One of the benefits of using a two-way cross generic area is reliable estimates could be created using 3 years of data instead of 5 years as mentioned in **Table 1**. When accounting for non-urban portion of the county, these new GAE should be lower than the current three-way cross generic areas that assumes completely urban areas. To determine if these new GAE are any different than SAE, we would use statistical tests to see if these new GAE for counties are statistically different than the ones produced with SAE. If most of the tests are non-significantly different then these new GAE would be a great tool for data users without access to Census facilities to produce reasonable subnational estimates at the county level with relatively available information to the public.

Another method is to use the three-way cross generic areas Census region by urbanicity by population size and used the method describe above to create estimates for each combination. The disadvantage to this approach is that using 5 years worth of data is required and, most likely, involves using the Census's NCVS datasets to figure out the proportion of a county that is urban for the 3 way generic areas cross.

Both generic areas and SAE are two methods provided by BJS for subnational estimation on the NCVS. While both methods have their advantages and disadvantages, they are great tools for data users wanting to understand local crime patterns. When determining which approach is best for an analysis, analysts should consider the advantages and disadvantages detailed in this paper to determine which method is most appropriate.

Acknowledgement

The authors would like to thank BJS for sponsoring this research and comments provided by Lynn Langton. However, 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, Michael (2012, 10-12 January). *Approaches to Estimating Subnational Victimization Estimates*. Paper presented at the Federal Committee on Statistical Methodology, Washington, D.C. Retrieved from https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/05/Planty_2012FCSM_I-B.pdf

Shook-Sa, B., Lee, P., & Berzofsky, M. (2015). *Assessing the Coverage and Reliability of Subnational Geographic Identifiers in the NCVS Public-Use File*. Retrieved from http://www.bjs.gov/index.cfm?ty=pbdetail&iid=5520

Fay, R. & Diallo, M. (2015). *Developmental Estimates of Subnational Crime Rates Based on the National Crime Victimization Survey*. Retrieved from http://www.bjs.gov/index.cfm?ty=pbdetail&iid=5499