Preliminary Effects of Oversampling on the National Crime Victimization Survey

Katrina Washington, Barbara Blass and Karen King
U.S. Census Bureau, Washington D.C. 20233

Note: This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on methodological issues are those of the authors and not necessarily those of the U.S. Census Bureau.

Keywords: oversampling, crime, NCVS

I. Introduction

The National Crime Victimization Survey (NCVS) is a nationally administered survey sponsored by the United States Justice Department's Bureau of Justice Statistics and conducted by the United States Census Bureau. The data collected from the NCVS is used to provide a detailed picture of crime incidents, victims, and trends from the victim's perspective. Data are collected twice each year from a nationally representative sample of approximately 60,000 households. Self-reported data is collected from every person 12 years of age and older within each household once every six months.

Over the past few years, national crime rates have steadily declined. Criminal victimization is already considered a rarely occurring event. By 2001, only 12.5% of all households interviewed reported any type of victimization, compared with 17.3% in 1999. As time goes by, this number is sure to continue to decrease as the crime rates themselves continue to decline. With the decline in crime itself and the rising costs of survey operations, the sponsors of the NCVS are faced with increasing financial burdens in maintaining such a large survey. For this reason, the sponsors of the survey are always looking for ways to decrease costs or at least maintain costs at their current level. Because over 75% of the households interviewed and over 90% of the persons interviewed for the NCVS are already not reporting any type of criminal victimization, the sponsors of the survey are not eager to reduce the sample size of the survey as a cost reduction effort at this time. However, costs are a concern. For this reason, the sponsors are now interested in finding ways to capture more criminal victimizations while at least keeping costs at their current level. More criminal victimizations give more cases to study. This has led to the idea of oversampling as a method of capturing more crimes for the money.

If oversampling can be used in such a way as to capture more crimes while also decreasing survey costs, then that would be an added incentive. The sponsors of the survey are interested in oversampling in such a way that sample is reduced in areas of the country where crime rates are low and costs are high while simultaneously increasing sample in areas where crime rates are high and costs are lower. The sponsors have a theory that travel costs are lower in high urban areas and higher in more rural areas. These ideas have developed into research on modifying the design of the NCVS to include oversampling of households in high crime urban areas of the country. This paper presents the initial research done on the topic of oversampling as it pertains to the NCVS. All research and findings presented in this paper are based on several simulations using NCVS data from 1999, 2000, and 2001.

II. Background

The NCVS is an on-going address-based survey. Data for the NCVS are collected each month of the year. A full sample of approximately 60,000 households is interviewed each half of the year. The target population for the survey consists of all civilian, non-institutionalized persons 12 years of age or older. The sample design for the NCVS is a stratified, multi-stage clustered design. The first stage primary sampling units (PSUs) are counties, groups of counties, or large metropolitan areas. The second stage (within-PSU sampling) units are randomly assigned clusters (segments) of approximately four housing units or housing unit equivalents. The PSUs are divided into self-representing (SR) PSUs and non-self-representing (NSR) PSUs, with approximately two-thirds of the sample located in SR PSUs. SR PSUs usually include more urban and metropolitan areas while NSR PSUs usually represent the more rural areas of the country. Currently, there are 93 SR PSUs

and 110 NSR PSUs in sample for NCVS. As a mixed-mode survey, both Paper and Pencil Interviewing (PAPI) and Computer Assisted Telephone Interviewing (CATI) modes of data collection are used in the NCVS.

III. Methodology

The sponsors of the NCVS are interested in capturing more crimes while maintaining or decreasing collection costs, all without changing their target sample size. They are hoping that this effort will have the additional side benefit of improving the reliability of the crime estimates. One way that this goal may be achieved is by implementing an oversampling design for the NCVS that maintains the target sample size determined in the original design. In order to meet this goal, our research approaches oversampling as a type of sample reallocation. The objective is to reallocate sample from PSUs with low crime rates, based on the Federal Bureau of Investigation's (FBI) Uniform Crime Reporting (UCR) data, into PSUs with high crime rates, also based on UCR data. Since SR PSUs usually include more urban areas, where crime is traditionally more common, we want to reallocate sample from low crime PSUs into high crime SR PSUs. The research in this paper looks at two methods in which this objective can be met: one where sample is reallocated from low crime SR PSUs into high crime SR PSUs and one where sample is reallocated from low crime NSR PSUs into high crime SR PSUs. High crime and low crime PSUs were chosen independently for each year included in the analysis based on that year's UCR data.

We began with a methodology that would yield very conservative estimates. This methodology, although not used to generate the findings reported in this paper, is the basis for determining the number of cases that are eventually reallocated across PSUs. The goal of the method was to reassign ALL cases[1] from the bottom 10%[2] of PSUs (bottom being those with lowest UCR crime rates) to the top 10% of PSUs (top being those with highest UCR crime rates). Given 93 SR PSUs and 110 NSR PSUs in sample for NCVS, we reassigned all the cases in 9 SR PSUs and 11 NSR PSUs with low crime rates into 9 SR PSUs with high crime rates. The reassignments of the cases in the low crime NSR and SR PSUs were done separately. This method resulted in approximately 3800 cases in the low crime NSR PSUs reallocated to the high crime SR PSUs, and 3650 cases in the low crime SR PSUs reallocated to the high crime SR PSUs. This method in essence eliminates the entire sample from these low crime PSUs. Eliminating the entire sample from a PSU eliminates the PSU itself from the sample. This in turn diminishes or completely eliminates the respective field representatives' (FR) workloads. Because of this probable impact, we decided not to pursue this avenue further. However, we did opt to use the actual number of cases reallocated for both the SR and NSR PSUs in our current research approach. Therefore, our research will look at the effects of reallocating approximately 3800 cases from low crime NSR PSUs to high crime SR PSUs as well as the effects of reallocating approximately 3650 cases from low crime SR PSUs to high crime SR PSUs. The data files used in our research each contain one year's worth of cases for the NCVS, which includes two full samples of approximately 60,000 households each.

In order to oversample cases in the SR PSUs with high UCR crime rates without increasing the target sample size determined in the original design, we reallocated a percentage of the sample from PSUs with low crime rates to the PSUs with higher crime rates. For low crime PSUs, a maximum of 25% of cases in each of these PSUs are dropped out of the sample. The remaining cases in these low crime PSUs are then weighted up by an *adjustment factor* so that the overall weighted number of households for each PSU remains unchanged.

To reallocate the sample just dropped from the low crime PSUs to the high crime PSUs, up to 50% of the cases in each high crime PSU are duplicated (one duplication per case) so that the number of cases dropped from the low crime PSUs equals the number of cases duplicated in the high crime PSUs. Of course, the number of cases dropped and the number of cases duplicated are not always exactly equal due to the initial sizes of the PSUs. For each PSU in which "new (duplicated) cases" are created, all the cases in the PSU are downweighted by an *adjustment factor* so that the overall weighted number of households for each PSU remains unchanged. This method allows us to retain (as close as possible) the weighted total number of households in sample for each PSU. Since our research involves looking at the possibility of reallocating cases from low crime NSR PSUs as well as low crime SR PSUs, this process is performed twice; once for low crime SR PSUs and once for low crime NSR PSUs. Table 1 shows the number of cases that

[1] *Cases* refer to households.

[2] 10% is an arbitrary amount of sample used. May be changed for future research.

were dropped and duplicated as well as the original total number of cases in SR and NSR PSUs for each year included in the study.

There were no distinctions made between interviewed and noninterviewed households when dropping and/or duplicating cases so there is no guarantee that all cases reallocated to high crime PSUs will become actual interviewed cases. Similarly, there were no distinctions made between households with crimes or without crimes so there is no guarantee that all cases reallocated will have a crime to add to our total. PSUs are classified as low crime or high crime based on their rankings for property crime rates, personal crime rates, and a cumulative personal and property crime rate formed from data contained on the UCR. Cases are chosen for deletion and duplication within the respective PSUs randomly using simple random sampling in the SURVEY SELECT Procedure of SAS.

The *adjustment factor* that is used to downweight or upweight the cases in PSUs where households have been dropped or duplicated is calculated as follows:

$$\textit{Adjustment Factor} = \frac{\textit{original PSU sum}}{\textit{new PSU sum}}$$

where

original PSU sum = the sum of the final household weights for the entire PSU, before any cases were deleted or duplicated,

and

new PSU sum = the sum of the final household weights for the entire PSU, after all cases have been removed or added.

The *adjustment factor* is then multiplied by the final household weight for each individual household in the PSU, after all cases have been dropped or duplicated, to produce a new final household weight for each household in the affected PSU. No additional reweighting is done.

Once all the adjusting, deleting, duplicating, and reweighting is done, the revised datasets are run through an NCVS variance program which produces new counts of weighted and unweighted crimes, new crime rate estimates, coefficients of variation (CVs), and standard errors. Changes in the CVs and the actual crime rates are calculated to detect any effects from the reallocation. For the standard error of computed crime rate differences, a correlation of one is assumed since the samples and the crimes reported are theoretically the same before and after reallocation. The gains in the standard error of the difference due to using such a correlation would come from the overlapping portion of sample only, which accounts for roughly 96% of the original sample.[3]

For us to say that one allocation method is more successful than the other, we will be looking at the following measures:

1. an increase in the number of unweighted crimes,
2. no significant differences in the crime rates,
3. a decreasing trend in standard errors, and
4. a decreasing trend in coefficients of variation.

The findings from the analyses are reported in the next section.

IV. Results

Results of the analysis for each reallocation method and for each year can be found in Tables 2, 3, and 4. Table 2 shows the unweighted number of personal and property crimes before and after the reallocation simulations. Table 3 shows the crime rates before and after the reallocation simulations. All rates have been rounded to the nearest tenth. Table 4 shows the percent change between the original crime rates (or CVs) and the crime rates (or CVs) after the simulated reallocation scenarios. All percent changes have been rounded to the nearest hundredth. Data is shown for only the major crime categories, where rates tend to be more stable over time. It is unfortunate that we were not able to obtain collection cost data to measure how each reallocation method might impact costs. We will assume for this analysis that costs remain at their current level for both methods.

A. Personal Crimes

There are few differences seen between the personal crime results for the two reallocation methods so they will be discussed together. Disappointingly, both reallocation methods have shown a decrease

[3] $\mathrm{Var}(x-y) = \mathrm{Var}(x) + \mathrm{Var}(y) - 2PR\,SE(x)SE(y)$ where P is the percent of overlap or 0.96 and R is the correlation of 1.0. See Kish, "Survey Sampling", page 459.

in the unweighted number of personal crimes. The negative percent change of personal crimes, for some years, approaches the percent of sample reallocated or 3.5%. For both methods, the weighted crime rates show distinct patterns by year. For 1999, we see a decreasing pattern in the crime rates, which is similar to the unweighted results mentioned above. On the other hand, we see what appears to be an increasing crime rate pattern for calendar year 2000 which is the opposite to what the unweighted count of personal crimes is telling us. Results based on the Wilcoxon Sign-rank test show the pattern for both years are statistically significant[4]. Finally for calendar year 2001, there is a mixed pattern. In other words, there are a balanced number of increases and decreases in the crime rates. This result is attributed to random effects.

The differences in the crimes rates were statistically significant for almost every personal crime category for all three years. Finally, the majority of the standard errors and the CVs appear to have an increasing pattern with both reallocations methods. It can be deduced that the increases in standard errors must outweigh the changes in the crime rates, thus causing the CVs to increase, although for most personal crime rates the changes appear small.

B. Property Crimes

Overall we see an increase in the unweighted number of property crimes for both reallocation methods. The greater increases are seen in the NSR PSUs to SR PSUs reallocation method. For this reallocation method, the positive percent change ranged from 0.35 to 1.47. For both methods of reallocation and for each year, we see a significant increasing pattern in the crime rates according to the Wilcoxon Sign-rank test, which is consistent with the unweighted results mentioned above. Almost all of the differences in the crime rates for the individual property crime categories are statistically significant in the positive direction.

The majority of the standard errors appear to increase with both methods. This is surprising since we have more crimes to work with. There were more distinct differences seen between the two reallocation methods when it came to the CVs of the property crimes rates. The SR PSUs to SR PSUs method seems to be producing a statistically significant increasing pattern in all property crimes according to the Wilcoxon Sign-rank test. The NSR PSUs to SR PSUs method is showing a mixed pattern of results by year. For the NSR PSUs to SR PSUs method, the changes in the standard errors overall were more consistent, in size, with changes in the crimes rates themselves. For instance, one of the largest percent increases in the crimes rates is in motor vehicle theft. However, the same crime category shows a decrease in its CV, which is highly desirable.

V. Summary

In summary, although only approximately 3.5% of all households in sample were chosen for reallocation, our research appears to show that oversampling for the NCVS, using either type of reallocation method studied, will have a significant effect on the crime rate estimates and the measured errors of those estimates. Discouragingly, many of the effects are not at all desirable.

Our first measure of success was whether we saw an increase in the number of unweighted crimes. We did see some consistent increases in the overall number of unweighted property crimes. The NSR PSUs to SR PSUs method seemed to give us the biggest and most consistent increases. Unfortunately, we didn't see an increase in the overall number of unweighted personal crimes.

The second measure of success was that we see no significant differences in the crime rates. Under both reallocation methods, for both personal and property crimes, we saw significant differences in the crime rates across almost all individual crime categories. It is possible that the large number of statistically significant differences across the crime rates is likely due to the large percentage of overlap in the samples.

It is necessary to note also that there is a pattern of increases in the personal crime rates under both methods of reallocation. This pattern of increase is a matter of concern because it seems to contradict what we were seeing in the unweighted totals. It implies that fewer personal crimes weighed more after reallocation.

[4] All significance is measured at the $\alpha = 0.10$ level.

The effect on standard errors and CVs are the third and fourth measures under evaluation. We saw increases in the standard errors, as well as in some of the CVs instead of the decreases desired. For personal crimes, this might be understandable since we have fewer crimes to work with. We are confused by this result for the property crimes with their increase in the number of crimes overall. The NSR PSUs to SR PSUs method seems to be most promising by showing a mixed pattern of increase and decreases in its percent change in CVs.

Our interest in doing this research is more than finding a method of oversampling that will increase the number of unweighted crimes, but one that will give us the biggest increase and the smallest error measures possible. Of course, we would like to find a method of oversampling that will produce the most uniform results across all crime categories, both personal and property. However, this may not be possible.

Given our objectives stated above, it is our preliminary recommendation that further research needs to be conducted before pursuing oversampling as an alternative survey design for the NCVS. However, if an oversampling design is adopted based on the results of the research presented in this paper, then we recommend that the reallocation method that moves cases from low crime NSR PSUs to high crime SR PSUs be researched more thoroughly as a possible oversampling design for the NCVS.

VI. Suggestions for Future Research

More research is definitely needed before any type of oversampling is introduced into the sample design of NCVS.

Future researchers need to look at how we assigned the high-crime/low-crime ranking to the PSUs. Perhaps the way we defined our ranking criteria for this analysis depended more on property crime characteristics than ones for personal crimes. Personal crimes are more rare than property crimes. Future researchers may want to look into applying larger weights to the personal crimes as a means of compensation, or possibly even look into finding crime information at a lower level of geography. Also, if implemented, this sample design will most likely be in place for the next decade. For this reason, and due to the variability of the UCR data for any given year, aggregate UCR data should be used to make final determinations of PSUs identified as high crime and low crime.

The weighting adjustment factor also needs to be reviewed. It may be better to apply it to the base weight and then perform NCVS's regular weighting and estimation routine on the reallocated samples which would take into consideration the change in the ratio of interviews to noninterviews for each simulation.

Additional thought also needs to go into how the reallocation impacts the clustering of the sample and its impact on our balanced half-sample replication system that we use to calculate variances. The concern there is that the reallocation may be inflating the clustering effect in some areas and throwing the replicates off-balance.

Only a small number of simulations were performed for the research outlined in this paper, with no substantial differences observed in the results. Future researchers may want to perform a large number of simulations, suggestively 100, and then observe the variability of the results produced from the simulations to determine if the actual cases chosen for deletion or duplication significantly impact the outcomes observed.

Of course, as stated in the introduction, the idea of oversampling as an alternative sample design for the NCVS is being investigated as a possible method to cut costs in fielding this survey on a continual basis. It would be useful to include interviewing costs at the PSU level in the next phase of this research. This design change will only be beneficial if the reduction in costs due to reducing sample in the NSR PSUs exceeds the cost of adding sample in the SR PSUs.

Additionally, different percentages of dropped and duplicated cases should be reviewed. It may be possible that larger savings can be experienced without compromising the integrity of the estimates, if, say, 40% of cases from low crime PSUs were dropped instead of 25%. In the same manner, more specialized research will need to be done to determine the optimal *number* of cases to be dropped and duplicated so as to receive the maximum benefit from the oversampling design.

Finally, once the oversampling design is implemented, a controlled experiment will need to be conducted to ascertain the actual impact that the design has on the estimates and the standard errors and CVs once in place. Remember that the research presented in this paper is based solely on a finite number of *simulations* of oversampling. We cannot fully anticipate the problems and pitfalls that an

oversampling design may encounter once fielded in real time.

References

Monahan, M.M. (2001), Memorandum for cc List, "National Crime Victimization Survey (NCVS) Monthly Survey Report : Quarter 1 through Quarter 4, 2000", Demographic Surveys Division, U.S. Census Bureau.

Monahan, M.M. (2002), Memorandum for cc List, "National Crime Victimization Survey (NCVS) Monthly Survey Report December 2001", Demographic Surveys Division, U.S. Census Bureau.

Victimization and Expenditure Branch (1997), Memorandum for Documentation, "Formulas for Standard Error and SIGMA Programs based on Generalized Variance Function Parameters: Use with the National Crime Victimization Survey", Demographic Statistical Methods Division, U.S. Census Bureau.

"Criminal Victimization in the United States, 1993", (1996) Bureau of Justice Statistics.

Table 1: Number of Cases Reallocated by Method Used

YEAR	1999		2000		2001	
Original Number of Cases in SR	68115		68805		69534	
Original Number of Cases in	38501		38878		39560	
Reallocation Method Used	NSR to	SR to	NSR to	SR to	NSR to	SR to
Number of Cases Dropped from	3847	3683	3705	3810	3658	3804
Number of Cases Duplicated in	3854	3667	3708	3800	3660	3827

Table 2. Unweighted Number of Crimes by Method Reallocation

Overall Personal Crimes				Diff. of NSR to SR versus Original		Diff of SR to SR versus Original	
	Original	NSR to SR	SR to SR	Number	Percent Change	Number	Percent Change
1999	2467	2409	2390	- 58	- 2.35	- 77	- 3.12
2000	2168	2120	2127	- 48	- 2.21	- 41	- 1.89
2001	1946	1934	1908	- 12	- 0.62	- 38	- 1.95
Overall Property Crimes				Diff. of NSR to SR versus Original		Diff of SR to SR versus Original	
	Original	NSR to SR	SR to SR	Number	Percent Change	Number	Percent Change
1999	8459	8515	8410	56	0.66	- 49	- 0.57
2000	7711	7739	7719	28	0.36	8	0.10
2001	7298	7406	7310	108	1.47	12	0.16

Table 3: Crimes Rates (and Standard Errors) by Method of Reallocation

	1999			2000			2001		
Type of Crime	Original Crime Rate	Crime Rate for NSR to SR Allocation	Crime Rate for SR to SR Allocation	Original Crime Rate	Crime Rate for NSR to SR Allocation	Crime Rate for SR to SR Allocation	Original Crime Rate	Crime Rate for NSR to SR Allocation	Crime Rate for SR to SR Allocation
Personal Crimes	33.7 (1.15)	33.4 (1.16)	33.4 (1.17)	29.1 (1.12)	29.2 (1.14)	29.4 (1.16)	25.9 (0.95)	26.0 (0.96)	26.0 (0.97)
Crimes of Violence	32.8 (1.14)	32.4 (1.15)	32.5 (1.16)	27.9 (1.09)	28.0 (1.11)	28.2 (1.13)	25.1 (0.94)	25.2 (0.95)	25.2 (0.96)
Robbery	3.6 (0.25)	3.4 (0.24)	3.4 (0.25)	3.2 (0.24)	3.3 (0.25)	3.3 (0.25)	2.8 (0.23)	2.7 (0.24)	2.6 (0.23)
Assault	27.4 (1.07)	27.3 (1.08)	27.3 (1.09)	23.5 (0.98)	23.6 (0.99)	23.7 (1.00)	21.2 (0.86)	21.4 (0.87)	21.4 (0.88)
Aggravated Assault	6.7 (0.39)	6.6 (0.40)	6.6 (0.40)	5.7 (0.35)	5.7 (0.36)	5.8 (0.36)	5.3 (0.36)	5.3 (0.37)	5.4 (0.37)
Simple Assault	20.8 (0.84)	20.7 (0.87)	20.7 (0.86)	17.8 (0.78)	17.9 (0.81)	18.0 (0.81)	15.9 (0.71)	16.1 (0.73)	16.0 (0.72)
Property Crimes	198.0 (3.77)	200.3 (3.89)	199.0 (3.87)	178.1 (3.90)	181.0 (4.05)	181.3 (3.99)	166.9 (3.95)	169.2 (4.01)	169.2 (4.06)
Burglary	34.1 (1.23)	34.5 (1.24)	34.5 (1.29)	31.8 (1.13)	32.3 (1.11)	32.5 (1.17)	28.7 (1.07)	29.0 (1.11)	28.8 (1.08)
Motor Vehicle Theft	10.0 (0.54)	10.1 (0.55)	9.7 (0.53)	8.6 (0.44)	8.9 (0.45)	8.8 (0.45)	9.2 (0.52)	9.2 (0.51)	9.4 (0.53)
Theft	153.9 (3.06)	155.6 (3.23)	154.8 (3.15)	137.7 (3.26)	139.8 (3.48)	140.0 (3.32)	129.0 (3.37)	130.9 (3.44)	130.9 (3.46)

Table 4: Percent Change in Crime Rates and Coefficients of Variation by Method of Reallocation

	1999				2000				2001			
	Percent Change				Percent Change				Percent Change			
Type of Crime	In Rates (NSR to SR)	In Rates (SR to SR)	In CVs (NSR to SR)	In CVs (SR to SR)	In Rates (NSR to SR)	In Rates (SR to SR)	In CVs (NSR to SR)	In CVs (SR to SR)	In Rates (NSR to SR)	In Rates (SR to SR)	In CVs (NSR to SR)	In CVs (SR to SR)
Personal Crimes	-0.93	-0.86	2.11	2.62	0.45	1.01	1.23	2.58	0.40	0.36	0.95	2.07
Crimes of Violence	-0.98	-0.90	1.98	2.57	0.57	1.06	1.15	2.39	0.51	0.40	1.04	1.99
Robbery	-4.74	-5.43	1.95	4.23	1.64	2.31	3.08	3.10	-2.06	-3.84	2.73	3.70
Assault	-0.45	-0.53	1.94	2.61	0.42	0.89	1.28	2.27	1.03	0.83	0.22	1.51
Aggravated Assault	-0.88	-0.97	4.43	3.51	-0.40	1.04	3.78	1.32	-0.36	1.42	2.68	1.13
Simple Assault	-0.32	-0.39	3.46	2.31	0.69	0.85	3.20	3.06	1.50	0.62	1.81	0.81
Property Crimes	1.15	0.52	2.14	2.26	1.66	1.78	2.14	0.46	1.35	1.34	0.06	1.32
Burglary	1.31	1.22	-0.21	3.72	1.59	2.13	-3.45	1.80	1.28	0.48	2.04	0.64
Motor Vehicle Theft	1.32	-2.37	0.69	1.63	3.45	1.34	-1.64	1.09	0.32	2.45	-1.17	0.93
Theft	1.11	0.56	4.46	2.49	1.56	1.73	5.38	0.34	1.44	1.45	0.42	1.14