Methods to Impute Household Income in the National Crime Victimization Survey

Andrew Moore[1], Marcus Berzofsky[1], Darryl Creel[1]
Tommy Holder[1]
[1]RTI International, 3040 Cornwallis Rd, Research Triangle Park, NC 27709

Abstract
This paper presents an analysis to determine the best approach to impute household income in a large, national survey. The National Crime Victimization Survey (NCVS) is a multi-stage, rotating panel design of households sponsored by the Bureau of Justice Statistics. The NCVS is designed to allow estimation of annual counts and rates of criminal victimization for both the population as a whole as well as subgroups of interest. Due to the strong relationship between socioeconomic status and criminal victimization, household income is a key characteristic often chosen to partition the population. Like many other surveys, the NCVS suffers from a high rate of missing data on household income and, while weighting is used to adjust for unit nonresponse, nothing is currently done to address item nonresponse. Failure to properly account for this missing data can lead to a loss of power and potentially biased estimates. We evaluate several potential approaches to imputing missing income data in the NCVS and assess each option on several criteria including consistency, variability, and usability. Final imputed results are also compared to the ACS for external validation of the chosen method.

Key Words: Imputation, Nonresponse, National Crime Victimization Survey (NCVS)

1. Introduction

With almost any large-scale survey, analysts will encounter incomplete data in some form. This challenge can present itself in two broad categories: (1) unit nonresponse and (2) item nonresponse (Lohr, 2010). Unit nonresponse occurs when a sample member fails to provide any information, either because they refused to participate in the survey or because they could not be contacted, or when the sample member fails to provide enough information for their record to be considered complete. Item nonresponse occurs when a unit respondent is unable to or refuses to provide information for one or more of the questions asked in the survey. In longitudinal or panel surveys, a special type of unit nonresponse that can occur is wave nonresponse. Wave nonresponse occurs when a sample member responds in at least one wave of data collection but fails to respond in another wave.

Typically, weight adjustments are used by most surveys to account for unit and wave nonresponse. This process helps ensure that the responding sample members are representative of the population of interest and is a crucial step in reducing the bias of point estimates. While most surveys employ some form of weight calibration to counteract the effects of unit and/or wave nonresponse, the handling of item nonresponse can vary greatly from survey to survey or even item to item within a survey. This inconsistency in the treatment of item nonresponse stems from several competing factors and interests. However, if not properly addressed, the survey can suffer from a loss of power and potentially biased estimates as a result of item nonresponse.

The most common approach for dealing with item nonresponse is statistical imputation in which missing data is replaced with valid response codes. However, the imputation process can be both time consuming and costly. Development and implementation of a system to perform statistical imputation could potentially delay the release of results and data to the public and/or negatively impact other aspects of the survey given a fixed amount of available resources. Therefore, agencies must carefully balance these interests while concurrently considering the primary rationale of the survey: producing relevant, accurate, and precise estimates of some characteristic of the population. Another factor that must be considered when deciding whether or not to address item nonresponse is the amount of data that is missing for a given item as well as its utility to analysts. Variables with high rates of missing data increase the potential for producing biased results but if the variable is of little benefit to researchers then item nonresponse may be less problematic. Conversely, even low rates of item nonresponse on a key outcome measure or domain variable could be enough of a concern to warrant imputation. The most problematic situation arises when a particular variable is subject to high rates of nonresponse while also providing considerable analytic utility. However, even when the decision is made to address item nonresponse through imputation, the selection of an appropriate method can be difficult. Numerous methods of varying complexity have been developed to perform imputation yet there is no consensus on the most appropriate method for all situations. As with the decision on whether or not to perform imputation to address item nonresponse, the decision on which method to implement must also carefully weigh various aspects of the particular situation along with the implications of that decision.

In this paper we compare two broad approaches to imputation, a linear model approach and a hot-deck approach, with each method being implemented under a single imputation and a multiple imputation framework. The basis for this comparison is household income as measured in the National Crime Victimization Survey (NCVS) with each method being evaluated on several criteria to assess both the feasibility and validity of the given approach. While each situation is unique and these results are not necessarily generalizable to all instances, the goal is to present an outline of the various considerations that must be made in choosing an appropriate method to address item nonresponse and provide a general framework for comparing these different options given the limitations and objectives of the particular survey and variable(s) of interest.

1.1 Background

The National Crime Victimization Survey (NCVS) is one of the primary sources of information on criminal victimization in the United States. Fielded since 1973, the NCVS is a nationally representative multi-stage household survey aimed at collecting detailed information about the victims and consequences of crime. Each year approximately 90,000 households and 160,000 persons are interviewed on the frequency and characteristics of criminal victimization (Truman and Langton, 2015). The NCVS employs a rotating panel design in which households are interviewed at six month intervals over a three year period for a total of seven interviews. Each wave, all residents in a selected household 12 years of age or older are interviewed about personal victimizations they may have experienced in the previous six months. One individual from the household is also selected each wave to answer questions related to household characteristics and property crimes the household may have experienced over the same timeframe. Through the collection of data on the number and characteristics of victimizations experienced by respondents, the survey allows estimation of annual counts and rates of personal and household criminal victimization as well as the comparison of national crime statistics over time and demographic characteristics.

1.2 Measurement of Income in the NCVS

In the NCVS, household income is measured by a single question in which the household respondent is asked to report the sum of income received by all household members 14 years of age or older living in the sampled housing unit during the 12-month period immediately preceding the month in which the interview occurs. Typically, income is only measured during the first, third, fifth, and seventh interviews with a carry-forward imputation approach being used for waves in which income is not asked. For this question, the respondent is asked to choose from one of the 14 categories shown below in ***Table 1***.

Table 1: Income Response Categories in the National Crime Victimization Survey

Income Code	**Amount in Dollars**
1	< 5,000
2	5,000 – 7,499
3	7,500 – 9,999
4	10,000 – 12,499
5	12,500 – 14,999
6	15,000 – 17,499
7	17,500 – 19,999
8	20,000 – 24,999
9	25,000 – 29,999
10	30,000 – 34,999
11	35,000 – 39,999
12	40,000 – 49,999
13	50,000 – 74,999
14	75,000 or more

As with many other large surveys, the NCVS has suffered from high rates of item nonresponse on the measurement of income and the rate of missing data has increased over time. As shown in ***Figure 1***, the rate of item nonresponse for income in the NCVS has increased from less than 20 percent in 1999 to approximately 30 percent from 2007 – 2013. Similar rates of nonresponse for income are also observed in the American Community Survey (ACS) with 31.5% of values requiring imputation in 2013 (U.S. Census Bureau, 2013).

The relationship between criminal victimization and socioeconomic status is well documented with respondents living in poor households experiencing more than double the rate of violent victimization as persons living in high-income households (Harrell, et al., 2014). However, for most variables in the NCVS, including household income, nothing is currently done to address item nonresponse. Ignoring this source of error and formulating conclusions about the relationship between criminal victimization and income based only on respondents from households with non-missing values could lead to erroneous interpretations of this association. Therefore, given the high rate of missing data and the strong relationship between income and many key outcome measures in the NCVS, item nonresponse should be addressed through a formal imputation process to reduce the potential for biased results and false conclusions.

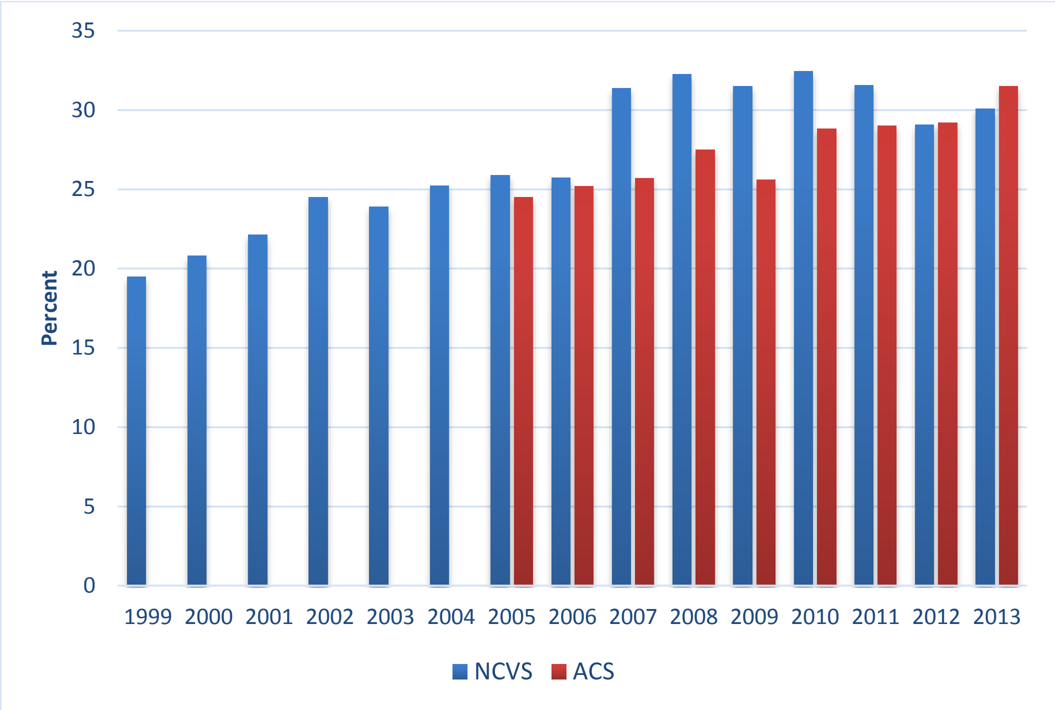

Figure 1: Rate of Income Item Nonresponse in the NCVS and American Community Survey (ACS), 1999 – 2013.

2. Imputation Methods

As previously illustrated, the choice of an appropriate imputation method for a given situation is not an easy decision. In the NCVS, this decision is further complicated by both the longitudinal nature of the survey and the planned missingness caused by income not being asked in every wave. This addition of intentional missingness leads to two nonresponse mechanisms: (1) missing at random for the planned missing values, and (2) missing not at random for the item nonrespondents. However, the panel design of the NCVS opens up more possibilities for dealing with nonresponse. Along with cross-sectional imputation methods such as mean or median value imputation, hot deck imputation, and regression based imputation in which only data available from the wave in which the item nonresponse occurs is utilized, longitudinal imputation methods are also available with panel data (Twisk & de Vente, 2002; Engels & Diehr, 2003). These longitudinal methods exploit information available from other waves and include methods such as carry forward imputation, carry backward imputation, mean or median value imputation based on the respondent's reported values in other waves, linear interpolation methods, and longitudinal regression methods where time and previous non-missing values are used along with other predictors. In most situations, longitudinal imputation methods are preferred over cross-sectional methods when dealing with panel data (Twisk & de Vente, 2002; Engels & Diehr, 2003; Tang et al. 2005; Watson & Starick, 2011).

Along with the choice of an imputation methodology, the analyst must also decide whether to perform single or multiple imputation. Multiple imputation methods are often preferred over single imputation methods because the standard errors more accurately reflect the variance due to imputation. However, multiply imputed data often increase the complexity of performing analyses and further burdens agencies and organizations when creating and releasing multiple versions of data files. For these reasons, the non-statistical considerations often outweigh the benefit obtained from performing multiple imputation with the recognition that variance may be under estimated.

For the current analysis, based on an understanding of the data, the level and mechanisms of nonresponse, and the advantages and disadvantages of the different imputation approaches, two imputation methods were chosen as possible solutions for dealing with item nonresponse in the household income measure of the NCVS. The methods chosen for further analysis include an explicit model based method and a hot deck method with both options being implemented from a longitudinal perspective. For each method, the performance of a single imputation approach and a multiple imputation approach will also be analyzed.

2.1 Implementation of Chosen Imputation Methods

The two methods chosen for consideration, hot deck and linear model, along with the single imputation and multiple imputation frameworks create four different options to analyze.

- SI-HD: The single imputation hot deck approach randomly chooses a donor from among a pool of respondents and assign's the recipient, the unit respondent with a missing income value, the donor's response.

- MI-HD: The multiple imputation hot deck approach is similar to the SI-HD approach with the process being repeated a specified number of times.

- SI-LM: The single imputation linear model approach uses a model-based approach to predict the missing income value for a recipient using a set of auxiliary variables related to income.

- MI-LM: The multiple imputation linear model approach is similar to the SI-LM approach with the process being repeated a specified number of times.

The linear model approaches were implemented using the Imputation and Variance Estimation Software, IVEware, from the University of Michigan's Survey Research Center (Raghunathan et al., 2002). This software implements an iterative process of sequential regression models and can be conducted with or without multiple imputation (Raghunathan et al., 2001). The imputation process proceeds sequentially through each of the interview waves and imputes missing income values using a linear model that includes non-missing income values from other waves, either imputed or provided by the respondent, and other predictor variables included in the model. Once all missing income values have been imputed, the process starts over and additional cycles are performed for a specified number of iterations. Given the ordinal nature of the income variable, the most appropriate choice of model would be a cumulative logistic regression model. However, this option is not currently available in IVEware so the variable must be treated as a nominal categorical variable or as a continuous variable. For the purpose of the current analysis, the decision was made to treat the ordinal income variable as continuous and use rounding to produce valid integer values.

The hot deck method was implemented in a two-step approach. In the first step, the tree package in R was used to identify key predictors of household income through recursive partitioning. This process evaluates the amount of variability in a group, or node, of observations and determines if splitting the group by another covariate would decrease this variability by a specified amount. The splitting of nodes continues until a terminal node is reached and the addition of more variables no longer decreases the variability enough to warrant the creation of additional groups. Thirty-seven variables were identified as potential predictor variables of household income. However, after running the tree package, the only variables used to create splits were income values from other waves. Upon completion of this stage, the second step commenced by utilizing the terminal nodes from the tree package as imputation classes. Once the imputation classes were formed, a weighted sequential hot deck (WSHD) was performed using SUDAAN's HOTDECK procedure (Cox, 1980; Iannacchione, 1982; Research Triangle Institute, 2012). As with the linear model approach, a cycling process was also implemented with imputed income values being updated during each subsequent cycle.

For the methods that included multiple imputation, the number of imputations performed was varied due to the high rate of item nonresponse in the income variable. By varying the number of imputations, the level of variability in the imputation process itself could be analyzed. The number of imputations performed for each method included 5, 10, 15, 20, and 25 imputations.

2.2 Advantages and Disadvantages of Imputation Approaches

Each of the four options under consideration have both advantages and disadvantages as presented in ***Table 2*** that must be included in the decision making process. The single imputation approaches have the advantage of producing a single analysis dataset which can be easier to analyze. This is particularly true for a survey such as the NCVS where the data are already divided into a hierarchical structure to represent the household level, person level, and incident level. In general, single imputation methods are also easier and less time consuming/costly to implement. While often more complex, the multiple imputation methods are able to account for the variability due to imputation which helps prevent analysts from underestimating the variance of point estimates. This additional variability is typically ignored with a single imputation approach where imputed values are treated as known rather than the result of a random process. An advantage of the two hot deck approaches is that they always produce valid categorical values which is not always the case with model-based methods. However, the linear model approach has the advantage of allowing the analyst to incorporate many auxiliary variables into the imputation model which could provide for a better estimate of the missing income value. With a hot deck approach, the number of auxiliary variables that can be used is often limited by constraints on the size of donor pools.

Table 2: Advantages of Different Imputation Approaches

Imputation Approach	Ease of Implementation	Produces valid categorical values	Single version of dataset required for analysis	Ability to incorporate many auxiliary variables	Accounts for variability in the imputation process
SI-HD	X	X	X		
MI-HD		X			X
SI-LM	X		X	X	
MI-LM				X	X

2.3 Assessment Criteria

To determine the most appropriate methodology for imputing household income in the NCVS, each of the four imputation methods were assessed based on three criteria. These criteria included the following:

1. Consistency of point estimates: how consistent is the distribution of imputed values when compared to the respondent data and/or an external data source?

2. Variability of the imputations: how much variation in the imputed values does the imputation procedure create?

3. Usability and ease of implementation: how easily can an analyst use the imputed data in an analysis?

3. Results

3.1 Consistency of Point Estimates

The first criteria used to assess the different options is the consistency of point estimates produced by each imputation procedure. To evaluate this, the distribution of household income for each method is presented in ***Figure 2*** along with the distribution of the respondent-only data. The distributions for the four imputation methods are each based on both respondent and imputed data. Based on this analysis the following findings are noted:

- The income distributions for the two hot deck approaches (SI-HD, MI-HD) are most similar to the respondent-only data.

- Under the multiple imputation framework (MI-HD, MI-LM), the number of imputations performed had little effect on the distribution of income. In Figure 2, only the distributions based on 5 imputations are presented as the results based on 10, 15, 20, and 25 imputations were nearly identical.

- Very little difference is observed between the single imputation approaches (SI-HD, SI-LM) and their multiple imputation counterparts (MI-HD, MI-LM).

- The linear model approaches (SI-LM, MI-LM) tend to impute more nonrespondents to the middle income categories (5 – 12) while fewer values are imputed for the low or high income categories. This tendency flattens out the income distribution and marks a distinction between the linear model and hot deck approaches. For instance, in the respondent-only data, 28.8% of households reported an income of $75,000 or more while only 5.0% of households with missing income were imputed to this category under the SI-LM approach. In contrast, 30.2% of households with missing income were imputed to this category under the SI-HD approach.

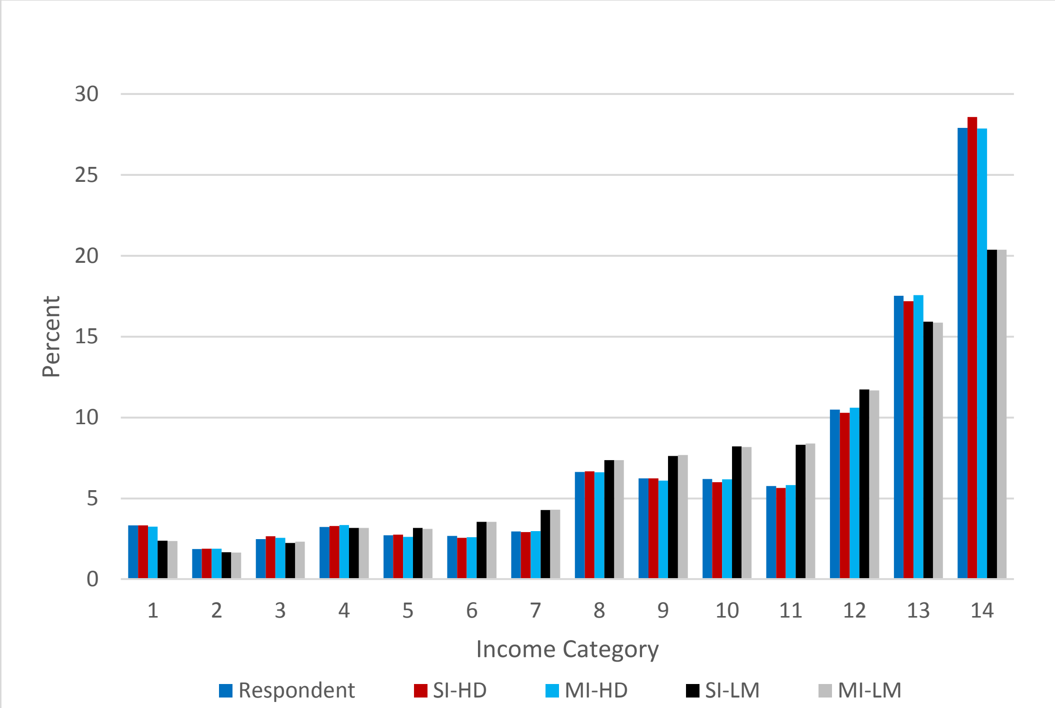

Figure 2: Distribution of Income by Imputation Method in 2010, Quarters 1 and 2.

3.1.1 Accuracy of Results

As illustrated in ***Figure 2*** and as previously discussed, the hot deck and linear model approaches produced different distributions of household income. While the hot deck approaches produced results that were most similar to the respondent data, because we don't know if the respondents and nonrespondents are similar with respect to household income, we are unable to definitively determine which method produces the more accurate results. To aid in this determination, two approaches were utilized.

1. The first approach compares the results from each method to an external estimate of the distribution of household income. Estimates from the American Community Survey (ACS) provide an appropriate benchmark for comparison as this survey produces annual estimates with high precision using similar income categories as the NCVS.

2. The second approach utilizes a Monte Carlo simulation to generate a population with known parameters (i.e., household income) that can then have missing values induced by different nonresponse mechanisms. These missing values are then imputed under each method and results are compared to the population with known values.

For the first approach, the distribution of income from the respondent only data and from each single imputation method, including both imputed and respondent data, were compared to estimates from the ACS as shown in ***Figure 3***. From this analysis we can see that the estimates produced using the hot deck approach are more similar to the distribution of income from the ACS than estimates produced using the linear model approach which tends to overestimate the percentage of households in the middle income categories while underestimating the percentage of households in the low or high income categories.

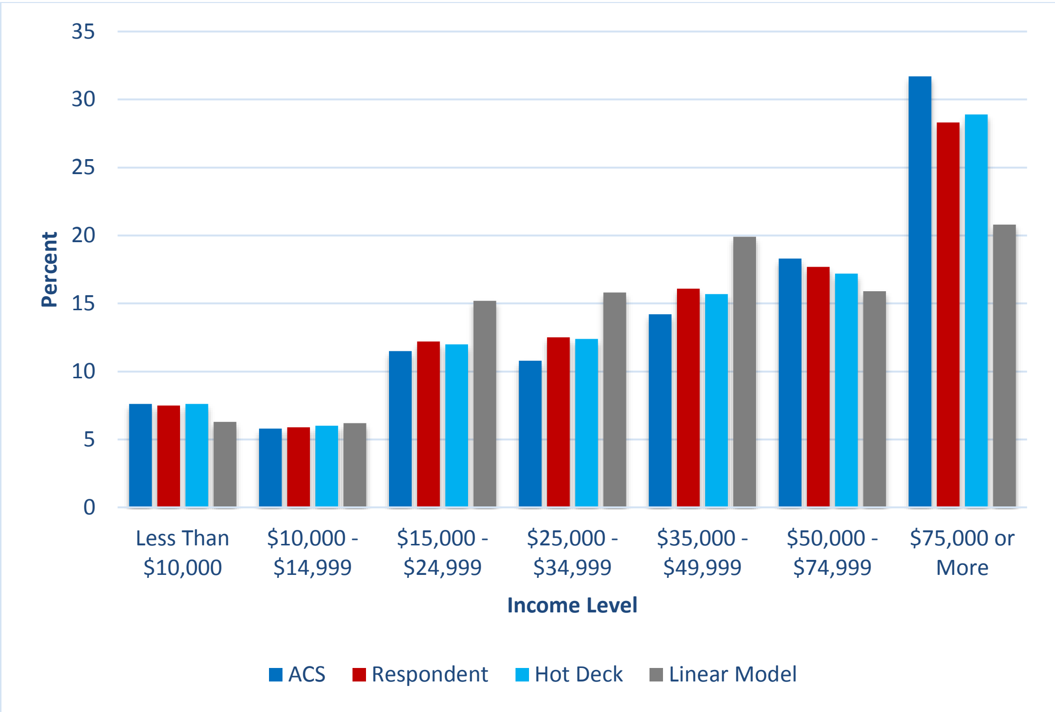

Figure 3: Comparison of Income Distribution for Respondent and Imputed Data from the National Crime Victimization Survey and the American Community Survey (ACS), 2010.

In the second approach to assessing the accuracy of imputation methods, a Monte Carlo simulation was conducted with the different imputation approaches being judged according to three primary criteria: absolute bias, relative bias, and confidence interval coverage of the population value. Analysis of the results from this simulation led to the following findings:

- The multiple imputation hot deck (MI-HD) approach performed best with respect to confidence interval coverage of the population value.

- The single imputation hot deck (SI-HD) approach performed well with respect to bias and relative bias.

- The linear model approach (SI-LM, MI-LM) performed relatively poorly with respect to coverage, bias, and relative bias.

3.2 Variability of Imputations

The second criteria used to assess the different imputation options was the precision with which each method produces estimates. Given that the NCVS is an annual survey, whichever imputation method is chosen must be able to produce estimates consistently to ensure that year-to-year trends are not unduly impacted. In ***Figure 4***, the standard error of the percentage of households in each income category is plotted by imputation method. As evidenced in this graphic, the single imputation approaches reduced the variability of estimates relative to the respondent-only data and, as expected, were also smaller than their multiple imputation counterparts. The hot deck imputation methods generally had larger standard errors than the linear model approaches. Also, the number of imputations performed for each multiple imputation approach had little impact on the standard errors. For this reason, ***Figure 4*** only includes the standard errors based on 5 multiple imputations for the MI-HD and MI-LM approaches.

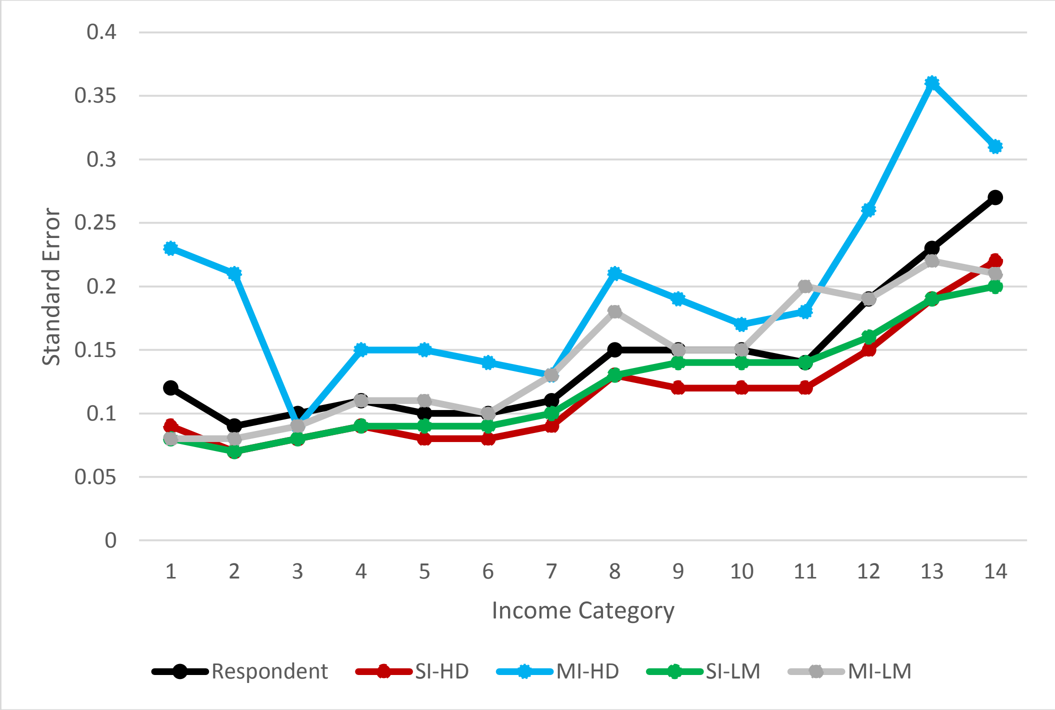

Figure 4: Standard Errors by Income Category and Imputation Method in 2010, Quarters 1 and 2.

A second approach was also implemented to assess the variability by which each method produces estimates. ***Figure 5*** presents the ratio of the multiple imputation standard errors to the single imputation standard error for both the hot deck and linear model methods by income category. Since the number of multiple imputations performed included 5, 10, 15, 20, and 25 imputations, each method should include five points for each income level. However, because the number of imputations performed did not have a large impact on the standard errors, not all points are discernable. The primary findings based on this display can be summarized by the following points:

- All ratios are greater than one indicating that the single imputation methods are likely underestimating the standard errors by treating the imputed values as known quantities without error.

- The standard errors from the multiple imputation hot deck (MI-HD) approach had more variability than the standard errors from the multiple imputation linear model (MI-LM) approach as evidenced by the wider range of points.

- The difference between the single imputation standard error and the multiple imputation standard errors were also of greater magnitude for the hot deck method.

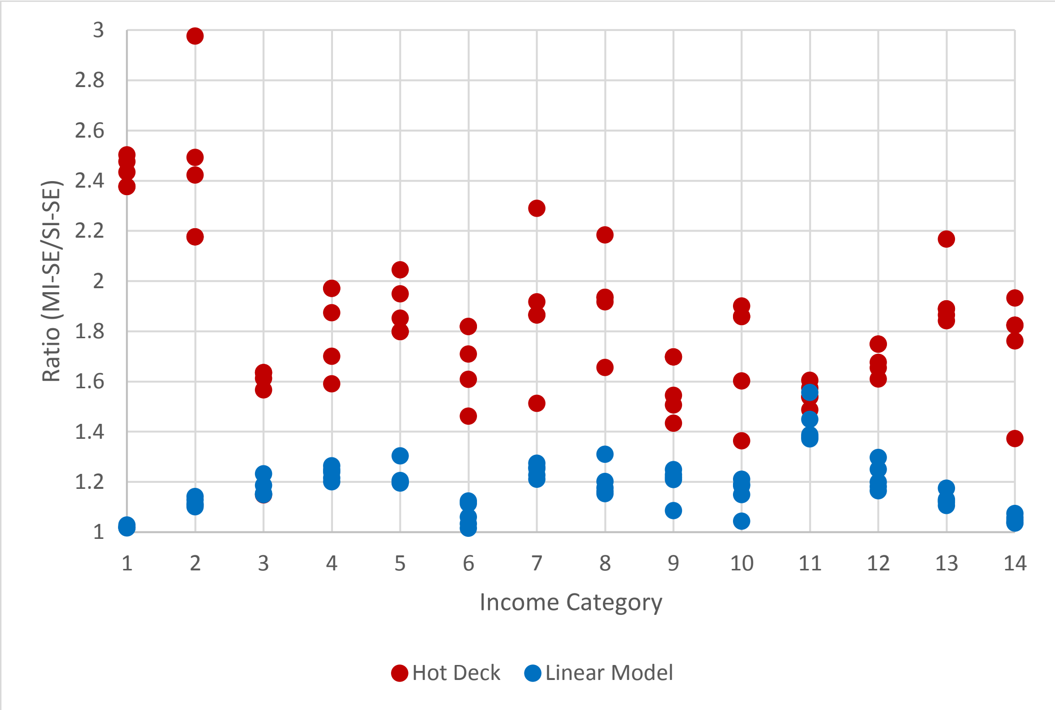

Figure 5: Ratio of Multiple Imputation Standard Errors to Single Imputation Standard Error by Income Category in 2010, Quarters 1 and 2.

3.2 Ease of Implementation

As previously discussed, imputation methods utilizing a hot deck approach are generally easier to implement than imputation methods based on a linear model. While the linear model approach does allow more covariates to be used as auxiliary variables, this feature also increases the complexity and requires the utilization of additional resources to deal with model fitting, convergence issues, and potentially model over-specification. In contrast, as implemented in the current approach, the hot deck method utilized the tree package in R to automatically identify the variables and levels required to form imputation classes with minimal input from the user other than a list of potential predictors. The hot deck method also ensures that imputed values are always integers for categorical variables such as income and reduces the potential for rounding error. Also, single imputation approaches are typically easier to analyze than multiple imputation approaches. While many commercially available software packages are now able to analyze multiply imputed data, such analyses increase the burden on end users particularly when it comes to managing multiple data files. This burden is further increased on a survey such as the NCVS which already includes multiple data files to represent the three levels of response: household, person, and incident.

4. Conclusions and Recommendations

Based on the findings presented above, the following conclusions and recommendations were made for imputing household income in the National Crime Victimization Survey:

1. Both the Monte Carlo simulation and the comparison of income distribution to an external data source, the American Community Survey, demonstrated that the hot deck method of imputation produced more accurate results than the linear model approach. Therefore, we would recommend using a hot deck approach for income imputation in the NCVS.

2. Despite the likely underestimation of standard errors, a single imputation approach is recommended for the imputation of income in the NCVS. This recommendation is made on the basis that very little difference was observed between the estimates produced by the multiple imputation approach and the single imputation approach and the single imputation approaches are typically easier to analyze and implement.

To validate these recommendations, the chosen method, SI-HD, was used to impute income in the NCVS for additional years (2008, 2009, and 2011) and the results were compared to the income distribution from the ACS. Those results are presented in ***Table 3*** along with the distribution from 2010 which was used for the analyses described previously.

Table 3: Distribution of Imputed Income in the NCVS and ACS, 2008 – 2011.

	2008		2009		2010		2011	
Income Category	**NCVS**	**ACS**	**NCVS**	**ACS**	**NCVS**	**ACS**	**NCVS**	**ACS**
Less than $10,000	7.5%	7.2%	7.3%	7.8%	7.6%	7.6%	7.8%	7.8%
$10,000 - $14,999	5.7	5.4	5.8	5.7	6.0	5.8	6.1	5.8
$15,000 - $24,999	11.9	10.7	11.9	11.2	12.0	11.5	12.2	11.4
$25,000 - $34,999	12.0	10.4	12.7	10.7	12.4	10.8	12.5	10.6
$35,000 - $49,999	16.4	14.2	16.8	14.4	15.7	14.2	15.7	13.9
$50,000 - $74,999	17.4	18.8	17.4	18.3	17.2	18.3	17.3	18.0
$75,000 or more	29.1	33.4	28.1	31.7	28.9	31.7	28.5	32.5

Acknowledgements

The authors would like to thank BJS for sponsoring this research and for the comments provided by Mike Planty and 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.

The work for this paper was completed using Federal funds provided by the U.S. Department of Justice under Cooperative Agreement number 2011-NV-CX-K068. As part of that work, a more detailed technical report was prepared by RTI International for the Bureau of Justice Statistics and is available for download at the following link: https://www.ncjrs.gov/pdffiles1/bjs/grants/248563.pdf

References

Cox, B. G. (1980). The weighted sequential hot deck imputation procedure. In Proceedings of the Survey Research Methods Section (pp. 721–726). Alexandria, VA: American Statistical Association.

Engels, J. M., & Diehr, P. (2003). Imputation of missing longitudinal data: A comparison of methods. *Journal of Clinical Epidemiology*, 56, 968-976.

Harrell, E., Langton, L., Berzofsky, M., Couzens, G. L., & Smiley-McDonald, H. (2014). *Household poverty and nonfatal violent victimization, 2008 – 2012.* Bureau of Justice Statistics Special Report (NCJ 248384). Washington, DC: U.S. Department of Justice, Office of Justice Programs. http://www.bjs.gov/content/pub/pdf/hpnvv0812.pdf

Iannacchione, V. G. (1982). Weighted sequential hot deck imputation macros. In Proceedings of the Seventh Annual SUGI—SAS Users Group International Conference (pp. 759–763). Cary, NC: SAS Institute.

Lohr, S. L. (2010). Sampling: Design and Analysis, Second Edition. Boston, MA: Brooks/Cole.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology, 27(1), 85–95.

Raghunathan, T. E., Solenberger, Peter W., & Van Hoewyk, J. (2002). IVEware: Imputation and variance estimation software, user guide. Ann Arbor, MI: University of Michigan, Institute for Social Research, Survey Research Center, Survey Methodology Program.

Research Triangle Institute. (2012). SUDAAN language manual (vols. I–II, release 11). Research Triangle Park, NC.

Tang, L., Song, J., Belin, T. R., & Unützer, J. (2005). A comparison of imputation methods in a longitudinal randomized clinical trial. Statistics in Medicine, 24, 2111–2128.

Truman, J. L., and Langton, L. (2014). *Criminal Victimization, 2013*, Government Printing Office, U.S. Bureau of Justice Statistics, Washington, DC.

Twisk, J., & de Vente, W. (2002). Attrition in longitudinal studies: How to deal with missing data. *Journal of Clinical Epidemiology*, 55, 329-337.

U.S. Census Bureau. (2013). *American Fact Finder: Table S1901: Income in the past 12 months (in 2013 inflation adjusted dollars)*. Retrieved from http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_1YR_S1901&prodType=table

Watson, N., & Starick, R. (2011). Evaluation of alternative income imputation methods for a longitudinal survey. Journal of Official Statistics, 27, 693–715.