Treatment of Missing Data in the FBI's National Incident Based Reporting System: A Case Study in the Bakken Region

Dan Liao, Marcus Berzofsky, David Heller, Kelle Barrick, Matthew DeMichele
RTI International, 3040 Cornwallis Road, Research Triangle Park, North Carolina 27709-2194

Abstract

In response to the growing reliance upon administrative records to generate national estimates of key indicators of interest, federal statistical agencies have been expanding and enhancing activities to assess and improve the quality of data collected through administrative records systems. Given that administrative records are often collected across different agencies or local reporting units, encountering missing data at both the individual and aggregated levels is inevitable and must be addressed when developing estimates. The Federal Bureau of Investigation's (FBI) National Incident-Based Reporting System (NIBRS) is a system designed to collect data from administrative records to be used for research and statistical purposes. It was developed as an expansion to its Uniform Crime Reporting (UCR) Program to improve the quality of crime data collected by law enforcement by capturing detailed information on each single crime occurrence. In this paper, we present an imputation method developed to handle missing data in NIBRS by leveraging other relevant external data sources. Given the hierarchical structure of NIBRS, this particular method addresses missing data occurring at multiple levels, including: a) incident level, due to item missing within each incident; b) agency level by month, due to some agencies reporting only for a partial year; and c) agency level by year, due to some agencies not submitting data to NIBRS for an entire year. The proposed method will be applied to a study in the Bakken region of the United States that utilizes the NIBRS data to examine how crime and law enforcement changed in the region as oil production increased from 2006 to 2012. A variance estimation method was also developed to evaluate the uncertainties in our estimates introduced by this imputation technique. In a broad sense, this research also can be viewed as an example of how to handle missing data in hierarchically structured administrative records.

Key Words: multilevel data; multiple imputation; item nonresponse; unit nonresponse; hot deck imputation

1. Background

1.1 UCR's SRS and NIBRS data

The FBI's Uniform Crime Reporting (UCR) Program is a voluntary data collection program which obtains information on offenses known to law enforcement and arrests from over 18,000 city, university and college, county, state, tribal, and federal law enforcement agencies annually. Since 1929, the UCR Program has collected information about crimes

known to law enforcement and arrests on seven main offenses. Each month, the traditional UCR Summary Reporting System (SRS) collects counts of the number of these crimes known to law enforcement.[1]

By the 1980s, criticisms of the UCR Program were commonly heard from law enforcement agencies, researchers, government policy makers, and the media. Many thought that the system needed to be expanded to cover a wider range of offense types and provide more detailed information on the nature of criminal incidents. The FBI responded to these criticisms by developing a more comprehensive and detailed reporting system. This system, which first began collecting data in the early 1990s, is known as the National Incident-Based Reporting System (NIBRS).

NIBRS is an incident-based reporting system for crimes known to the police. For each crime incident coming to the attention of law enforcement, a variety of data are collected about the incident. These data include the nature and types of specific offenses in the incident, characteristics of the victim(s) and offender(s), the location of the incident, types and value of property stolen and recovered, and characteristics of persons arrested in connection with a crime incident. As of May 2014, 32 states have been certified to report NIBRS to the FBI,[2] and three additional states (Georgia, Kentucky, and Mississippi) and the District of Columbia have individual agencies submitting NIBRS data. Fifteen states are 100% NIBRS reporters, meaning that all (or nearly all) of law enforcement agencies in the state submit only incident-based data to the NIBRS.

With the large amount of incident-based information that NIBRS gives researchers, policy makers, and governors, a much fuller understanding of crime can be gained. However, similar to other sources of administrative records, NIBRS data is plagued by missing data, which can cause significant bias in statistical estimation and obstructs analysts' ability to make inferences directly from the data. Therefore, proper treatment is desired to address the problem of missing data in analysis of NIBRS data. In this paper, we propose an imputation method to deal with missing data in NIBRS and apply it to a study that utilizes NIBRS data to examine how crime changed in the Bakken region of the United States as oil production increased from 2006 to 2012. Details on this study will be given in the following section.

1.2 Crime in the Bakken Region

U.S. crude-oil production grew by more than one million barrels a day in 2012, the largest increase in the world and in U.S. history (BP, 2013). The Bakken formation underlying North Dakota (ND) and Montana (MT) is a major oil boom site, with experts predicting as many as 48,000 new wells in the next 20–30 years. Men have migrated to the area around the Bakken formation in record numbers to work; ND's burgeoning temporary workforce lodging facilities (sometimes referred to as "crew camps") had a capacity of 37,000 in 2013.[3] Observers speculate that the influx of temporary workers has contributed to an increase in crime and public disorder in the small towns and rural areas that surround the drilling.

[1] SRS also applies a hierarchy rule. Generally speaking, only the most serious crime that occurred during an incident is recorded in the data. There are some exceptions to this rule, but they are rare. For example, homicide and arson are counted without regard to the hierarchy rule.
[2] See: http://www.jrsa.org/ibrrc/background-status/nibrs-states.html.
[3] This number is founded through the 2013 ND GIS Hub Data Portal at: http://www.nd.gov/gis/.

The purpose of this study is to use NIBRS and other available data sources to examine how crimes reported to the police, law enforcement responses to crime (arrests and clearances), and law enforcement staffing have changed in the Williston Basin/Bakken region as oil and natural gas production increased. In addition, trends in crimes will be compared between Bakken and non-Bakken regions in ND, MT, and South Dakota (SD). Although all three states (ND, MT, and SD) are 100% NIBRS reporters, missing data not only occurs at the incident level with item responses (e.g., victim-offender relationship), but also at the agency level with unit nonresponse. For unit nonresponse, some agencies did not report to NIBRS in some months of a year or even did not report in an entire year.

2. External Data Sources

External data sources used in this study include the UCR's SRS data, the UCR's Law Enforcement Officers Killed and Assaulted (LEOKA) data, the Bureau of Justice Statistics' (BJS's) law enforcement agency crosswalk (LEAIC) file, and the Census's annual population estimates data.

2.1 UCR's SRS Data

The UCR's SRS data is the FBI's original system for recording crimes and has been in active since 1929. It has tracked data on seven crimes since the beginning: murder, robbery, rape, aggravated assault, burglary, theft, and vehicle theft, and it started reporting on arson in 1979. The UCR's NIBRS, developed later in early 1990s, collects more comprehensive information on crimes but has less coverage in the US. Some agencies only report to SRS but not to NIBRS.

Although the three states included in the Bakken study—North Dakota, Montana, and South Dakota—are considered "full reporters," not every agency reported to NIBRS. In order to start assessing the quality of the NIRBS data reported, we examined each state's NIBRS reporting response rate: defined as the percentage of agencies within a given state that report to the SRS for at least 6 months in a given year that also report to NIBRS at least 6 months in a given year (i.e., we divided the number of agencies reporting to NIBRS at least 6 months by the number of agencies reporting to SRS at least 6 months). ***Table 1*** displays the number of agencies across the three states contributing data to SRS and NIBRS as well as the annual NIBRS response rates during the study period from 2006 through 2012. Although the response rate increases over time, from 72% in 2006 to 76% in 2012, the rate was lower than expected as all three states included in this assessment are considered by the FBI to be "full reporters." Clearly, there are some agencies in these "full reporter" states that only report via SRS.

Table 1: Annual Counts of Agencies Reporting 6 Months or More to SRS and NIBRS: 2006–2012

	2006	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*
SRS agencies	360	361	362	362	359	352	353
NIBRS agencies	254	280	285	293	301	299	305
Response rate*	71.74%	77.09%	78.92%	80.89%	84.16%	85.28%	87.01%

*: number of agencies reporting to NIBRS 6+ months / number of agencies reporting to SRS 6+ months (excluding state, special, and tribal agencies)

The reported monthly crime counts in SRS data can be very helpful in the imputation procedure for the NIBRS data (see Section 3). It should be noted that if an agency reported to SRS but not to NIBRS in a certain month, its monthly total crime counts in SRS cannot be applied to NIBRS directly and considered as its monthly crime counts in NIBRS, because SRS and NIBRS collect different crime information. For instance, SRS tracks only 8 serious crime types, while NIBRS tracks 46 different crime types, which include the 8 crime types in SRS and other crime types that are not reported in SRS. SRS reports only the most serious crime committed in a single incident (e.g., if a murderer has raped his victim, only murder is reported). NIBRS requires officers to report multiple offenses, victims, and offenders.

2.2 UCR's LEOKA and BJS's LEAIC Data

The UCR Program also collects data on the incidents in which law enforcement officers were killed and assaulted. This set of data is referred as LEOKA and also contains Police Employee (PE) data. LEOKA data consist of some key characteristics of each agency including: number of male/female sworn and civilian personnel, population size, and whether or not the agency is located in a metropolitan statistical area. Information in LEOKA was merged with NIBRS data at the agency level to facilitate imputation at the agency level. Another agency level variable, agency type, was extracted from the BJS's LEAIC because it is not available in LEOKA.

2.3 Annual Population Estimates from the U.S. Census

The annual population estimates produced by the U.S. Census are also merged with NIBRS data at the county level. The Census produces annual population estimates at the national, state, and county levels and can be used as indicators of recent demographic changes (https://www.census.gov/popest/). These estimates are available for all counties in the Bakken region in the period of interest (i.e., 2006–2012) and are disaggregated by sex and age groups.[4] A hypothesis underlying imputation at the agency level is that the crime counts can be related to the population estimates, and this correlation is proved in the current NIBRS data. The population estimates are not only used in the imputation procedure, but also adopted in the descriptive analysis when we are curious about how the crime trends will look when controlling for the growth of the population sizes measured by these population estimates.

3. Methodology

NIBRS data are collected by police officers in law enforcement agencies throughout the country. The crime data are submitted to the FBI either through their state's UCR Program or (in rare cases) the agency submits their crime data directly to the FBI. To optimize all the information collected through NIRBS in our descriptive analysis, missing values in key variables will be filled in using a statistical imputation procedure. Given the hierarchical

[4] We also considered using data from the American Community Survey (ACS), which provides 1-year estimates for geographic areas with a population of at least 65,000 residents; 3-year estimates for geographic areas with a population of at least 20,000 residents; and 5-year estimates for smaller areas. Of the 33 counties in the Bakken region, 29 only have 5-year estimates and the remaining 4 have 3-year estimates. However, the single year estimates are more ideal for our study because we are interested in the trend of population change from 2006 to 2012 in the Bakken region. Moreover, 5-year estimates are not available until 2009 (data collection began in 2005 and 2009 was the first year for which 5 years of data was available). In light of these limitations, we decided to use the census annual population estimates rather than the ACS data for population estimates.

structure of NIBRS, an imputation method is developed to treat data missingness occurring at two difference levels—(1) victim level, due to item missingness within each incident, and (2) agency level, due to underreporting (unit nonresponse) by some agencies. NIBRS and other external data sources were utilized through the imputation and estimation procedures.

3.1 Imputation at Victim Level

Missing data were first imputed at the victim level. There could be multiple victims in an incident. In the Bakken study, the characteristics of the victims and offenders are of interest in conjunction with crime rates and crime types (e.g., violent vs. property victimization). Therefore, the victim level is the lowest level in the NIBRS data and it could be more efficient to impute missing data at this level first without aggregating them into an upper level.

Item missingness at the victim level varies across the different variables. For example, some general demographic variables in NIBRS have very high response rates with less than 5% missingness, such as age, gender, and race, while other variables, such as gang involvement, have relatively low response rates. In this study, we focused on three variables with more than 5% missingness in one or more of the three states included in the study: (1) presence of weapon, (2) injury sustained, and (3) victim-offender relationship. These were imputed through a hot deck imputation procedure (Ford, 1983) using matching variables to identify a donor for each missing value.

The choice of matching variables were mainly based on the strength of their association with the variables to be imputed. First, all the external data sources described in Section 2 were merged to the NIBRS data at the victim level. Then, with collaboration of the subject-matter experts, we identified a list of candidate variables in the entire dataset that might be highly or moderately correlated with the variables to be imputed. On this list, we included some incident-specific variables, such as offense type, characteristics of the victim(s) and offender(s), and the location and time period of the incident. We also consider some agency-specific variables, such as agency type and the total number of officers in the agency. For each categorical variable on this list, two-way frequency tables were generated to examine the relationship between the candidate variable and each of the three variables to be imputed. For each continuous variable on this list, scatterplots were produced to examine the relationship between the candidate variable and each of the three variables to be imputed. We used scatterplots rather than regression models in case of outliers or nonlinear relationship. If a variable was shown in the frequency tables or scatterplots that it does not have a strong or moderate relationship with any of the three variables to be imputed, this variable will be dropped from the list. With all the variables remained on the list, regression models were applied further to select the matching variables that are mostly correlated with the three variables to be imputed. Note that the number of matching variables to be used is also limited in order to have adequate number of donors in most of the donor groups.

In the hot deck imputation procedure, the selected matching variables were ordered from the most important to the least important, denoted as "x(*1*), x(*2*)… x(*m-2*), x(*m-1*), x(*m*)," so that x(*1*) is the most important matching variable and x(*m*) is the least important. A group defined by these matching variables must meet two criteria to be eligible as a donor group: a) the number of donors in this group must be equal to or greater than 10 and b) the number of donors in this group must larger than the number of donees (i.e., cases with missing data) in the same group. The two criteria provide some warranty that the

imputation procedure is random and a donor won't be one of the few that are used repeatedly for a considerable number of donees. If a group fails one or two of the criteria, its neighboring groups are combined into a larger group so that the larger group is defined by variable "x(*1)*, x(*2)*… x(*m-2),* x(*m-1).*" If the larger group still fails one of the criteria, the next level of neighboring groups are combined into an even larger group until the resulting group meets both criteria. For example, for victim-offender relationship, there are 3,921 groups defined by the matching variables and 343 donor groups after group collapsing. Once the donor groups are all defined, a donor will be randomly drawn and its value will be assigned to a donee. This procedure is completed through PROC IMPUTE in SUDAAN 11 (RTI International, 2012). Matching variables used for each variable were listed in ***Table 2*** in the order that have been used in our imputation procedure.

Table 2: Matching Variables for Each Imputed Variable in the Order of Creating Donor Groups at the Victim Level

Order	*Weapon*	*Injury*	*Victim-Offender Relationship*
x(1)	Boom Period Onset	Boom Period Onset	Boom Period Onset
x(2)	State (1=MT, 2=ND/SD)	State (1=MT, 2=ND/SD)	State (1=MT, 2=ND/SD)
x(3)	Victim's Age	Victim's Age	Type of Victim (1=Individual; 2=Police Officer)
x(4)	Victim's Race	Victim's Race	Victim's Age
x(5)	Offender's Age	Offender's Age	Victim's Race
x(6)	Offender's Race	Offender's Race	Offender's Age
x(7)	Victim's Gender	Victim's Gender	Offender's Race
x(8)	Offender's Gender	Offender's Gender	Victim's Gender
x(9)	Victim's Ethnicity	Victim's Ethnicity	Offender's Gender
x(10)	--	--	Victim's Ethnicity

3.2 Imputation at Agency Level

After all of the data missing at the victim level is imputed, the resulting dataset has complete information for each incident reported by agencies. However, this does not address the problem of unit missingness due to nonresponse at the agency level. As mentioned in Section 2.1, some agencies did not report data for the full year or did not report at all during the year. In order to address this issue, we aggregated the NIBRS data at the agency level by month (referred as "monthly agency level"). To get aggregated information on victimization at the monthly agency level, we created a series of count variables at this level, such as the total number of violent victimizations, the total number of victimizations with female victims, and the total number of victimizations where a weapon was present. The count variables were used to calculate the overall yearly crime rates and yearly rates by male or female victims. This data was then merged with all the external data sources. In addition, SRS, LEOKA and LEAIC data were used to create a complete list of all the agencies[5] in MT, ND, and SD from 2006 through 2012. If an agency on this list is not listed in the NIBRS data at the agency level, this agency will be considered missing.

[5] We also shared this list with UCR program managers to ensure that we were not including agencies that had been disbanded, consolidated, or covered by another agency.

Because NIBRS is an incident-based system, an agency does not need to report to NIBRS if there is no incident in a month. However, the agency will report to SRS with a monthly crime count of zero. The first step in imputation at the monthly agency level is to distinguish the agencies who are truly missing with those who have zero crimes. If an agency's aggregated NIBRS count variables are missing in a particular month but it reported to a monthly crime count of zero to SRS, we will assign all its crime counts as zero for that month. After this step, if an agency still has missing counts, it will be considered to be truly missing and imputed through a hot deck imputation using matching variables to identify a donor agency.

The method used to select matching variables for donor imputation at the monthly agency level is similar to the method used at the victim level. A list of candidate variables were first identified and their relationships with the total number of violent and property victimizations in NIBRS at the monthly agency level were examined. Because some agencies did not report to NIBRS but reported to SRS (see ***Table 1***), we also examined the relationship between the candidate variables and the monthly crime counts in SRS. The candidate variables that are considered to be mostly correlated with the SRS's monthly crime counts and the total number of violent and property victimizations in NIBRS were used as matching variables.

The hot deck imputation procedure at the monthly agency level is similar to the one at the victim level. Matching variables were ordered from the most important to the least important and small groups were collapsed to make a donor group with adequate size. The one major difference was that the criteria used to collapse the groups was slightly different for the monthly agency imputation than for the victim level imputations. The first criterion was the same, specifically that the number of donors in this group must be equal to or greater than 10. However, the second criterion used was that the number of donors in this group must take over 25% of the total number of cases including donors and donees. We changed this criterion because with the data we have, there are only 11 donor groups if we use the criterion that the number of donors must be greater than the number of donees in the group. We relaxed the requirement here so that some of the key matching variables can be reserved in the definition of the donor groups, such as Bakken versus non-Bakken region, state, and agency type. Based on these revised criteria, we have 144 groups defined by the matching variables and 65 donor groups after group merging. Once the donor groups are all defined, a donor agency will be randomly drawn within each donor group, and the entire series of NIBRS count values from this donor agency will be assigned to a donee in the same group. This procedure is also completed through PROC IMPUTE in SUDAAN 11. The monthly crime counts of SRS were used in combination with the NIBRS data to identify matching variables at monthly agency level that are moderately correlated with the crime counts. Matching variables for imputation at monthly agency level are used in the order of: Boom Period Onset; State (MT, ND, SD); indicator of metropolitan statistical area; agency type; different population size groups; and agency groups with more or less male officers.

3.3 Multiple Imputation

Unlike a truly observed value, an imputed value is a statistical guess for a missing value with some uncertainties. In the hot deck imputation procedure, if we draw a donor repeatedly for a few times, we may end up with different donor in each time. To compensate and estimate the uncertainties introduced by imputation, instead of filling in a single value for each missing value, a multiple imputation procedure (Rubin, 1987) is used in this study.

In this procedure, we imputed each missing value multiple times so that they can represent the uncertainty about the right value to impute. First, we imputed the missing data at the victim level five times using the method described in Section 3.1 and created five separate datasets. Second, with each of the five data files, we created data at monthly agency level and imputed the missing data at the monthly agency level five times using the method described in Section 3.2. Therefore, we generated 25 (5 by 5) different imputed datasets at the monthly agency level.

For each outcome variable (e.g., total crime count), the mean of its estimates derived from the 25 imputed datasets is then used as the final estimate,

$$\bar{\theta} = \frac{1}{25}\sum_{i=1}^{5}\sum_{j=1}^{5}\hat{\theta}_{i(j)},$$

where *i* represents *i*th imputed dataset at the victim level, *i(j)* represents *j*th imputed dataset at the monthly agency level with *i*th imputed dataset at the victim level, $\hat{\theta}_{i(j)}$ is the estimate derived from the the *i(j)*th imputed dataset, and $\bar{\theta}$ is the final estimate.

With this multiple imputation method, the total variance estimator for 25 imputed dataset can be expressed as

$$T = \bar{U} + \left(1 + \frac{1}{25}\right)B,$$

where $\bar{U}$ is the average of the 25 imputed variances ("within imputation" component) and $B = (25-1)^{-1}\sum_{i=1}^{5}\sum_{j=1}^{5}(\hat{\theta}_{i(j)} - \bar{\theta})^2$ ("across imputation" component). From the design-based perspective, the imputed variance within each imputed dataset is equal to zero, because the dataset we are dealing with (i.e., NIBRS data) is from administrative records, which collect data from all the subjects in the population rather than a probability sample of the population. Therefore, we have $\bar{U} = 0$ and the total variance estimator based on the 25 imputed datasets in our case can be simplified as:

$$T = \left(1 + \frac{1}{25}\right)B$$

where $B = (25-1)^{-1}\sum_{i=1}^{5}\sum_{j=1}^{5}(\hat{\theta}_{i(j)} - \bar{\theta})^2$.

3.4 Evaluation Criteria

Coefficient of variation (CV) based on the total variance estimator is used to evaluate the uncertainties in the estimates introduced by our imputation method described in Sections 3.1 and 3.2. CV is a standardized measure of variation, which is defined as the ratio of the standard deviation to the mean $\sqrt{T}/\bar{\theta}$. The CV measures the relative amount of variability associated with the estimate. Low CV values indicate more reliable estimates. There are no steadfast rules as to what constitutes a reliable estimate. The reliability of an ACS estimate would be considered "good" if its CV is equal or less than 15%.[6]

4. Results

To evaluate our imputation method, three study variables were imputed, including presence of weapons, injury sustained, and victim-offender relationship. ***Table 3***, ***Table 4***, and ***Table 5*** present the annual final estimates and their CVs for these three variables in the MT, ND, and SD area. As can be seen from the tables, the CVs for most of the estimates are less than 1%, which indicates the uncertainties introduced by our imputation method

[6] See: http://www.ofm.wa.gov/pop/acs/userguide/ofm_acs_user_guide.pdf.

are relatively small in most of the cases. The largest CV is 4.01% for the estimate of violent victimization rate when the victim and offender are strangers. The value of this estimate is very small, 5.57 per 10,000 population. Similar to other statistical prediction methods, imputation method is usually less efficient and can cause more variation when the magnitude of an estimate is small. Overall, our imputation method performs well for these three variables with all of the CVs less than 5%.

Table 3: Presence of Weapons in Violent Crime Victimizations: 2006–2012 NIBRS*

Region	*Year*	*Rate (%)*	*CV (%)*
Bakken	2006	48.59	0.73
	2007	51.74	1.55
	2008	49.51	0.42
	2009	48.17	0.45
	2010	42.42	0.42
	2011	47.22	0.74
	2012	46.72	0.34
Non-Bakken	2006	46.40	0.18
	2007	46.63	0.27
	2008	47.79	0.23
	2009	49.17	0.24
	2010	48.81	0.23
	2011	48.98	0.20
	2012	48.98	0.11

*: Results here are preliminary and are subject to changes.

Table 4: Injury Sustained in Violent Crime Victimizations: 2006–2012 NIBRS*

Region	*Year*	*Rate (%)*	*CV (%)*
Bakken	2006	37.76	0.76
	2007	38.88	2.08
	2008	37.76	0.59
	2009	41.32	0.56
	2010	35.28	0.65
	2011	37.53	0.20
	2012	39.82	0.62
Non-Bakken	2006	23.46	0.33
	2007	24.86	0.28
	2008	25.88	0.34
	2009	28.45	0.47
	2010	27.56	0.26
	2011	28.19	0.13
	2012	26.76	0.15

*: Results here are preliminary and are subject to changes.

Table 5: Violent Victimization Rates (per 10k population), by Victim-Offender Relationship: 2006 and 2012 NIBRS*

Victim-Offender Relationship	*Year*	*Rate (per 10k population)*	*CV (%)*
Stranger	2006	5.57	4.01
	2012	6.93	1.28
Intimate Partner	2006	24.34	0.72
	2012	27.60	0.43
Acquaintance	2006	20.98	1.11
	2012	19.57	0.61
Family	2006	10.65	2.42
	2012	11.63	0.97
Other	2006	4.07	3.27
	2012	5.45	0.88
Unknown	2006	0.00	--
	2012	0.97	2.98

*: Results here are preliminary and are subject to changes.

5. Discussions and Future Research

5.1 Discussion

There are a few lessons we learned from this study. First, we found that proper strategies should be employed to deal with the large data volume of NIBRS. NIBRS is a nationwide incident-based reporting system that collects a variety of data about the incident. It is very challenging for analysts to process and analyze such a large amount of data, especially with a complex hierarchical structure. One possible solution is hardware that has increased memory and uses powerful parallel processing to process the data quickly. However, this solution can escalate the cost for data analysis, and sometimes analysts do not even have access to these advanced hardware.

The solution we adopted in this study is to subset the main data file and only extract useful variables for a particular analysis. For example, before the imputation procedure at victim level, we went through the NIBRS data file and all the other auxiliary data files and identified a list of potential predictors to be used in the imputation model, extracted them from the main data file, performed the imputation procedure within this subset and then merged the imputed variables back to the main data file. This procedure needs to be performed with caution. Analysts should gain adequate knowledge about the data and their research objectives to extract the right variables without missing any key information. Variable extraction could take a day or several days when the size of the main data file is extremely large. Therefore, it will be much more efficient for analysts to plan in advance rather than going back and forth with this procedure. Time is crucial for statistical tasks with such large data like NIBRS. Planning is very necessary to cut the cost and help analysts derive results in a timely manner.

Second, we recognized that data editing is essential before any imputation procedure when dealing with administrative records like NIBRS. Missing data can be caused by the setup of the system. When dealing with the injury sustained variable, we found that this variable is missing for all the murder incidence. This is because there is inherently no injury variable when homicides are coded in NIBRS. Recoding, rather than a random imputation procedure, should be applied in such cases. In our study, we assigned all the murder incidences as injury sustained before the imputation procedure.

5.2 Future Research

Some potential areas will be investigated in our future research. As we mentioned in Section 3.2, because NIBRS is an incident-based system, an agency will not report to NIBRS when no incident has occurred. Although we used SRS monthly crime counts to identify the zero-crime agencies at the monthly level, this method could be inaccurate and should be assessed further. SRS only tracks 8 serious crimes, while NIBRS tracks 46 different crimes. Even when an agency reported zero crime to SRS, this agency could have some incidents that should be reported to NIBRS. A more sophisticated statistical approach could be developed to make a better prediction on the zero-crime agencies at the monthly level.

In addition, the mechanism of missingness in NIBRS should be studied in depth. A thorough examination can be conducted to check the patterns of missing data at the victim level. In the example of missing injury data in the murder incidence, the missing data is caused by the setup of the system. If there are multiple victims in one incident, observed data in one victim record might be helpful to handle missing data in another victim record within the same incident. For missingness at the agency level, investigations, such as a short survey, could be made to follow up with agencies who did not report to NIBRS on monthly basis. In the system, we could add an option for agency to report zero crimes on monthly or yearly basis, so that researchers could easily distinguish the zero-crime agencies from missing agencies. Learning the mechanism of missingness in the data can help us to improve our current data collection and editing procedures. It can also facilitate the selection of good matching variables for donor imputation and the development of more sophisticate imputation methods.

Finally, the method described in this paper imputes the data at the victim level first and then agency level. It yields 25 imputed datasets when applying a multiple imputation procedure. Another method for this hierarchical data structure could be imputing the data at the agency level five times and then performing imputation at the victim level within each of the five datasets. The latter method will only generate five imputed datasets. However, in our method, imputation at the agency level is much easier because we only need to copy the summary record of the donor agency to the donee agency, while we need to copy all the incident records from the donor agency to the donee agency in the latter method. In addition, the imputation procedure at the victim level in the latter method could take much longer because the data volume could be increased significantly after the imputation at the agency level is completed. With our method, we can more easily estimate the variation caused by imputation at the victim level and variation caused by imputation at the agency level. With the latter approach, it is impossible to estimate variations caused by the two imputation steps separately. Further research can be conducted to evaluate the performance of both methods.

Acknowledgements

The work for this paper was funded by the Bureau of Justice Statistics Analytic Resource Center (BJS-ARC), Cooperative agreement (COA) 2012-R2-CX-K011.The authors would like to thank BJS for sponsoring this research and comments provided by Kimberly Martin and Alexia Cooper. 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

BP. (2012). *Statistical Review of World Energy*. London: British Petroleum Co.

Ford, B.L. (1983). An overview of hot-deck procedures. *Complete Data Sample Surveys, 2*, 185–207. Academic Press, Inc.

RTI International. (2012). *SUDAAN®, Release 11.0* [computer software]. Research Triangle Park, NC: RTI.

Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63*, 581–592.