Imputing National and Regional Crime: Innovative methods to imputing the National Incident-Based Reporting System (NIBRS)

Ashlin Oglesby-Neal, Dean Obermark, and KiDeuk Kim

Urban Institute. 2100 M Street NW, Washington, DC 20037

Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference

Introduction

Missing data is a widespread statistical problem for real-life datasets across disciplines. In this paper, we use machine learning algorithms to impute reported crime data. Specifically, we impute various annual crime category totals for agencies that do not report crime via the National Incident-Based Reporting System (NIBRS) and derive regional and national crime estimates for a broader group of crimes than is currently available. Our work shows that machine learning approaches are valuable alternatives to traditional imputation approaches.

Crime Data Collections

Crime is a focal point of public discourse. From the nightly news to local and national political campaigns, the United States' preoccupation with crime extends far and wide. Despite the seeming importance of crime to the general public, the nation's primary crime data collection program, Uniform Crime Reporting (UCR), is lacking. UCR has changed little since its creation in the late 1920s, and collects minimal information about a small set of crime types.

The UCR program provides a framework for standardized crime reporting and relies on the voluntary reporting of law enforcement agencies to the Federal Bureau of Investigation (FBI). The UCR program collects information on crimes and categorizes them into Part 1 and Part 2 offenses. Part 1 consists of offenses deemed to be more serious and less susceptible to fluctuations in reporting. As such, Part 1 crime is typically used to track changes in crime across time and place. The following offenses make up Part 1 crime: murder and nonnegligent homicide, rape, robbery, aggravated assault, burglary, motor vehicle theft, larceny-theft, and arson. Importantly, the UCR program applies a "hierarchy rule" to classifying criminal incidents in which multiple offenses occur, and only the most serious offense (according to the hierarchy) is reported. For example, in an incident where a person is robbed and then murdered, only the murder is counted in UCR. Though the hierarchy rule helps simplify reporting, it discards detail about additional offenses.

NIBRS is a more detailed and comprehensive crime data collection program that's intended to eventually replace UCR. NIBRS is more detailed relative to UCR in many ways. For example, NIBRS collects weapons information for all violent offenses, types of injury for each victim, date and time information, as well as incident location (e.g., indoor vs. outdoor) (Federal Bureau of Investigation 2016). These example details represent only a fraction of the 58 data elements collected per incident. NIBRS is more comprehensive than UCR in that it seeks to collect information about a much larger group of offense types than UCR. NIBRS uses a 24-offense category framework and collects information on 52 distinct "Group A" offenses. In addition, NIBRS does not employ a hierarchy rule, and can collect information on multiple offenses for each incident (ibid.). Though these data collection advantages promise a fuller view of reported crime, only approximately 37% of law enforcement agencies report crime via NIBRS, and agency reporting varies widely by state and region, hindering NIBRS utility in understanding crime levels (Federal Bureau of Investigation 2017).

Research Questions

If NIBRS were adopted more fully, it would paint a fuller picture of reported crime in the United States. NIBRS would illuminate criminal offense levels for a large body of crime that is not represented within UCR because of the

hierarchy rule and UCR's more limited offense categories. Moreover, NIBRS' expanded incident details provide important contextual information on crime and could allow for more accurate tracking of a bevy of more particular offenses as defined by incident characteristics (e.g., gun crimes, night crimes, etc.). We use machine learning to artificially extend NIBRS current data collection, and focus on two illustrative examples of how NIBRS comprehensiveness and detailedness bolsters our understanding of crime.

1. What are regional and national offense counts according to an imputed NIBRS dataset and how do these estimates compare with UCR?

2. What are regional and national counts of gun offenses according to an imputed NIBRS dataset and how do these estimates compare with UCR?

Method

Data

This analysis utilizes multiple publicly available data sets. Of the NIBRS data, we use the 2015 victim-level file, rather than the incident, offender, or arrestee-level files, because the victim-level file contains the most cases in its default structure. Using this file also allows us to estimate the full impact of crime on victims, as a single crime can have multiple victims. We use three other data sets provided by the Department of Justice: the 2012 Law Enforcement Agencies Identifiers Crosswalk (LEAIC), the 2015 UCR Program Offenses file, and the 2015 UCR Police Employee Data (LEOKA). The LEOKA and LEAIC contain agency-level information, such as the size and type of the agency. The UCR Offense file holds the level of recorded crime in each month reported per agency to the FBI. As we discussed above, the UCR Offense file provides information on a limited array of crime types, but is the only standardized way to examine crime across the United States.

Beyond crime and law enforcement agency data, we also use the 5-year American Community Survey (ACS), collected by the U.S. Census Bureau. This data set spans the years 2011-2016, and provides demographic information for counties across the country, including age and gender distribution, racial makeup, educational attainment, and homeownership composition, among many other covariates (approximately 2,000 in the file used). We pre-processed this data set using Principal Component Analysis to reduce the number of variables into a group of 185 principal components. This reduction in dimensionality makes applying the ML algorithms less computationally demanding.

Each of these files have different base units, necessitating the re-structuring/aggregation of certain files so that they can be merged with the others. Each case in the NIBRS victim-level file is a victim. We aggregate the total count of victims by each unique agency that participates in NIBRS (n = 6,278) per month, creating a file with 67,595 observations. Both the LEAIC and UCR data use reporting agency as the base unit, allowing us to easily merge them with the aggregated NIBRS data by the Originating Agency Identifier (ORI) code.

The ACS can be based in multiple geographical units. We use county as the base unit because each law enforcement agency exists in at least one county. Some law enforcement agencies cover more than one county, and if an agency does cover multiple counties, we pick the first county provided in the LEAIC. The first listed county should correspond to the county in which most an agency's jurisdiction lies, but we did not independently verify this. We then merge the county-level ACS data using the LEAIC file's county FIPS code. With this agency-level file constructed, we can impute estimates of NIBRS crime for each agency.

Imputation

We use machine learning (ML) to estimate two types of crime in NIBRS: total crime and crime involving guns. We choose these two outcomes because they present a unique set of challenges and difficulties for prediction. Regarding estimating total crime, the UCR provides estimates of total crime for the eight Part 1 offenses, and the number of total crimes reported in NIBRS is correlated to these counts, likely making imputation easier. For crime involving guns, in UCR agencies can report use of a gun for the four violent Part 1 offenses (murder, rape, robbery, and

aggravated assault), but theoretically guns can be used in a wider spectrum of offenses and NIBRS collects this additional information, likely making imputation harder.

We test several machine learning models to estimate these crime outcomes, including ensemble methods, neural network models, and support vector machines. We use multiple methods to see how reliably they predict crime levels, and to identify the most accurate method. For those interested in a technical treatment of different ML methods, we recommend James et al. (2013), which offers an excellent overview of machine learning.

Before testing different imputation methods, we divide our core analytic data file of 67,595 observations into three groups: training, test, and hold-out. The training set composes 50% of observations, while the other groups compose 25% of observations each. All models are developed on the training set, and the most accurate models are tested on the test set. The final model is validated on the hold-out set. This final model is then used to predict crime levels for the 15,846 agencies that do not report to NIBRS.

The success of imputation methods often lies in the strength of the predictors included in the models. Because of this, we include as many variables as computationally practical, and allow the machine learning methods to identify the most important predictors. We develop each model with the same predictors.

Validation: Sensitivity Check on NIBRS Missingness

For the validation of the final models, we test their performance by the percentage of agencies that report to NIBRS in each state. We explore the performance of the model in states for which 1-50% of agencies report to NIBRS, 51-75% of agencies report, and 76-100% of agencies report. We test the final models' performance for each of these groups because it is possible that the models would perform better in states with lower degrees of missingness and worse for states with higher degrees of missingness. All of the data processing and analyses are conducted in *R*.

Results

In the results section, we report the performance of the final models on the test and validation sets, and the estimated crime counts after imputation.

Final Models

The final models have a strong performance across several metrics on the test set, shown in Table 1. The best performing algorithm for estimating total crime was an Elastic-Net Regularized Generalized Linear Model (glmnet), with a R-squared of 0.995, a Root Mean Square Error of 26.28, and a Mean Absolute Error of 12.03. The mean predicted monthly crime is very similar to the mean observed crime (89.57 v 89.30). The median monthly predicted and observed crime are also very similar (21.41 v 21.00).

The best-performing model predicting gun-involved crime used Generalized Boosted Regression (gbm) with a Poisson distribution. The model has a R-squared of 0.943, a Root Mean Square Error of 6.19, and a Mean Absolute Error of 1.35. The mean predicted monthly gun-involved crime is very similar to the mean observed crime (2.89 v 3.00). The median monthly predicted and observed gun-involved crime are also very similar (0.50 v 0.00).

Table 1: Model Performance

	Total Crime	Gun-involved Crime
Model type	GLM Net	GBM Poisson
R^2	0.995	0.943
RMSE	26.28	6.19
MAE	12.03	1.35
Mean predicted	89.57	2.89
Mean observed	89.30	3.00
Median predicted	21.41	0.50
Median observed	21.00	0.00

The most important predictors in each model varied, and the top 10 are listed below. The most important ten predictors in the total crime model include characteristics of the agency and jurisdiction, as well as monthly crime counts from the UCR. In contrast the ten most important predictors in the gun-involved crime model are almost exclusively monthly crime counts from the UCR. One principal component, which is a summarized version of some of the ACS variables, was among the most important predictors for gun crime.

Total crime

1. Core city indicator
2. Manslaughter 2014
3. Sheriff's office indicator
4. State police indicator
5. Manslaughter 2015
6. Murder 2014
7. Attempted burglary 2014
8. Attempted rape 2015
9. Murder 2015
10. Other robbery 2015

Gun-involved crime

1. Gun agg. assault 2015
2. Gun robbery 2015
3. Gun agg. assault 2014
4. Hand/foot agg. assault 2014
5. Gun robbery 2014
6. Total crime 2015
7. Principal component 152
8. Agg. assault 2015
9. Knife assault 2015
10. Forcible rape 2015

Validation

In the validation, we evaluate the performance of the models on the holdout set, which is the remaining 25% of the sample. We separate these observations into three groups based on their state's level of participation in NIBRS (1-50%, 51-75%, 76-100%). The models have a strong performance across the groups, shown in Table 2. The total crime model performs well across all three groups, but the MAE is lowest for the group of agencies with high state-level participation in NIBRS. The performance of the gun-involved crime model varies more across the three groups. The R-squared is highest for agencies with low levels of participation in NIBRS, while the MAE is lowest for agencies with high levels of participation. Considering the models perform well and largely similar across all three groups, we feel comfortable using them to impute NIBRS estimates for non-reporting agencies.

Table 2: Validation Results

	Total Crime	Gun-involved Crime
Low participation		
R^2	0.99	0.99
MAE	14.7	1.61
Med participation		
R^2	0.99	.90
MAE	17.3	1.90
High participation		
R^2	0.99	0.94
MAE	11.3	1.32

Imputation

We then use these models to impute NIBRS estimates of monthly total crime and gun-involved crime for the 15,846 agencies that report to UCR, but not to NIBRS. We then aggregate these to be yearly estimates at the regional and national level, shown in Table 3. On average, the NIBRS estimate of total yearly crime is 1.67 times greater than the amount of crime reported in UCR Part 1. The NIBRS estimate of gun-involved crime is 2.16 times greater than the UCR gun-involved crime, which we define as homicide, gun robbery, and gun aggravated assault.