Assessing the Quality of Administrative Data for National Estimation of Crime Statistics: The National Incident-Based Reporting System

FedCASIC

April 5, 2022

Marcus Berzofsky, RTI International

Study Sponsors:

BUREAU OF JUSTICE STATISTICS - OFFICE OF JUSTICE PROGRAMS

FEDERAL BUREAU OF INVESTIGATION - CRIMINAL JUSTICE INFORMATION SERVICES

Motivation

- Many federal agencies are moving to administrative data as a supplement or replacement for traditional survey data based on a sample

- Administrative data is often collected from multiple sources (e.g., each state or some sort of establishment)

- The quality of the data provided is not always uniform and, unlike surveys, is not always as rigorously monitored and controlled

- How the data quality issues are addressed can impact the ability to use the administrative data for official statistics which require the data to be representative and of high quality

Types of Data Quality Issues for Administrative Data

Nonreporting. Nonreporting occurs when a reporting unit does not report any information during the year.

Partial reporting. Partial reporting occurs when a reporting unit only reports a subset of their information during the year.

Item-level missingness. Item-level missingness occurs when individual data elements from a submitted record are missing in a required field.

Unanticipated values. Unanticipated values are reported values that are technically valid per the reporting rules, but when looking across all reported data the values do not look reasonable.

UCR Background

- The Uniform Crime Report (UCR) – which provides information on the number of crimes reported to police – has been published by the FBI for almost 100 years
- Up until 1986, the FBI only collected data in ***summary*** form meaning law enforcement agencies (LEAs) provided only an aggregate count of the number of crimes which occurred each month by offense type
- Beginning in 1986, the FBI allowed LEAs to **submit data through summary form OR through the National Incident-Based Reporting System (NIBRS)**.
- NIBRS provides details about each specific incident
- **Beginning in 2021, the FBI allowed data to only be reported through NIBRS. This transition has an impact on the data quality of the UCR estimates.**

Nonreporting

Coverage Rate Over Time (National)
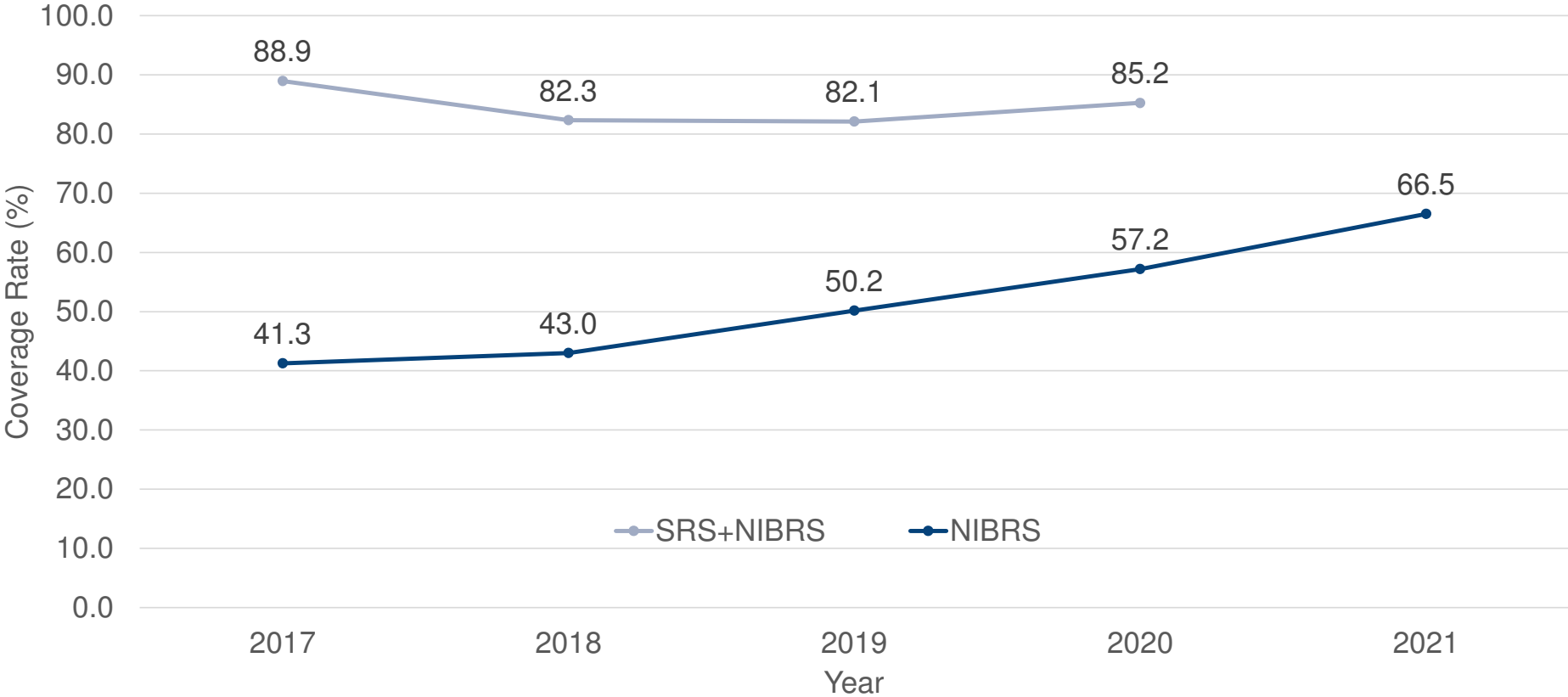
NOTE: Coverage is based on the number of reporting agencies

Mitigation Strategies for Nonreporting

1

Produce inferential weights to reduce bias

2

Produce confidence intervals based on mean squared error (MSE) to account for uncertainty in the estimates including the bias

3

Suppress estimates that are determined to be unreliable

Measuring Uncertainty: Using RMSE to Measure and Present Uncertainty

Measuring Variance	Use linearization method for estimating variance that can account for the calibration weighting procedure
Measuring Bias	Because a near census for some estimates were collected through 2020, prior summary+NIBRS data can be used as benchmark for estimating bias

$$MSE(X) = Var(X) + Bias(X)^2$$

$$RMSE(X) = \sqrt{MSE(X)}$$

- The RMSE can be used as the half-width of a confidence interval:

$$X \pm 1.96 \times RMSE(X)$$

- The estimation system will produce the confidence interval based on RMSE for all NIBRS estimates

Considerations for Suppression

Relative RMSE can be used to determine unreliable estimates

When the majority of estimates for a domain are suppressed then the entire domain may need to be suppressed

When coverage rates for a domain are extremely high then suppression should not be used regardless of relative RMSE

Partial Reporting

CONFIDENT

Types of Partial Reporting

01

Partial reporting: Agency reports less than 12 months of information

02

Outlier reporting: Agency reports for a month but the number of cases reported is an outlier relative to other months

Examples of Partial Reporting and Mitigation Methods

	Reporting Month											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Agency A	X	X	X	O	O	O	O	O	O	O	O	O
Agency B	O	O	O	O	O	O	O	O	X	X	X	X
Agency C	O	O	X	X	O	O	O	O	O	O	O	O

O=reported data in that month; X=did not reported data that month.

- **Mitigation Plan**: Block Imputation
- For each missing month, donor agency with similar characteristics and estimated crime counts identified
- Crime incidents from donor agency used for missing month

Example of Outlier Detection and Mitigation Strategy
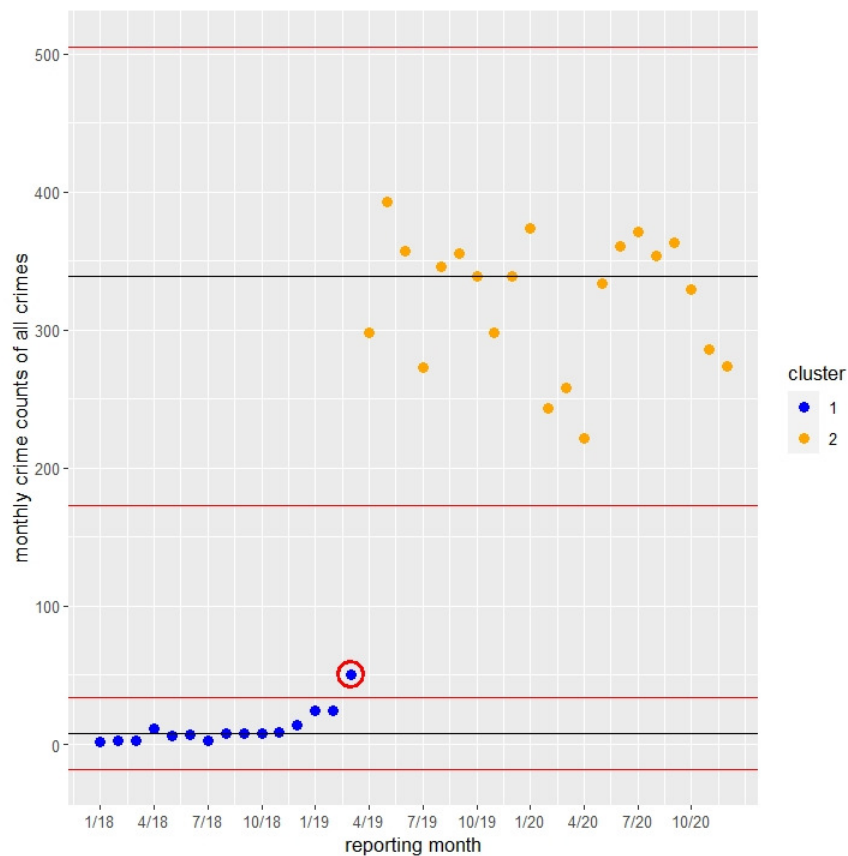
- **Mitigation Plan**: No direct mitigation is recommended
- Indirect mitigation of following-up with reporting agency (i.e., feedback loop) can be done to determine if outlier counts are correct

Unanticipated Values

CONFIDENT

Types of Unanticipated Values

High levels of "unknown" reports for a data element

High levels of particular response option relative to other agencies

Examples of Unanticipated Values and Mitigation Strategy

Incident time of day: The first response option is midnight. Some agencies have a disproportionally high number of crimes occur at midnight indicating the agency may have simply selected the first response option

- Mitigation Strategy: Feedback loop
- Identifying outlier agencies:
  - Agency has level of unanticipated values in the 95th percentile or outside two standard deviations
  - Distribution needs to be "wide" enough for identified agencies to be considered different from other agencies

Item Nonresponse

CONFIDENT

Types of Item Nonresponse

Items needed for estimation

Outcome items

Examples and Mitigation Strategies

- ***Items needed for estimation***: Rates are computed by age, race, and sex; therefore, missing values for these attributes among victims or arrestees are problematic

- ***Outcome items***: Victim-offender relationship is critical characteristic of a crime incident and used in many analyses

- **Mitigation Strategy**: For items needed for estimation, hot deck imputation used except for offender characteristics in "uncleared" (i.e., unsolved) cases

- **Mitigation Strategy**: For victim-offender relationship, hot deck imputation used except for offender characteristics in "uncleared" cases.

- For other data elements, no imputation conducted due to greater uncertainty in the imputation process

Summary

Summary

- Federal agencies are expanding their use of administrative data for official statistics
- High data quality is critical for ensure administrative data can produce representative estimates
- The methods used to ensure high quality data for NIBRS can be applied to other administrative data collections
- Breaking the types of data quality issues into 4 types is helpful in determining the best mitigation strategies
- The 4 types include
  - Nonreporting from the reporting units
  - Partial reporting from the reporting units
  - Item-level unanticipated values
  - Item-level nonresponse

Thank you

Contact: Marcus Berzofsky| email: berzofsky@rti.org