CJARS Data Quality Automation

Brian Miller

University of Michigan

October 26, 2022

Managing large-scale data with a small team

- CJARS collects data from hundreds of criminal justice agencies, from municipal police departments to county courts to departments of corrections

- More than 3 billion rows of raw data from 30 states covering 177 million criminal justice events and 38 million unique individuals

- Core goals of harmonization process:
  - Transform jurisdiction-specific data into a national schema
  - Maintain high level of data quality
  - Transparency – include both coded values and original text where possible

Outline

1. Overview of CJARS data processing pipeline

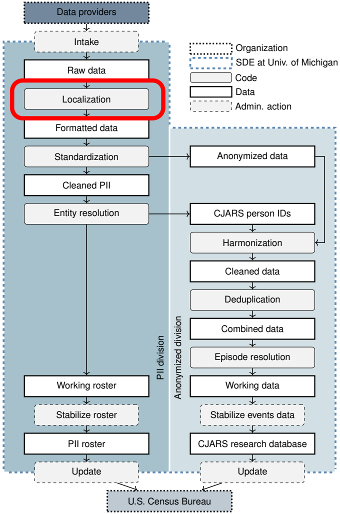
2. Four major processes for managing data quality
   - Real-time review
   - Benchmarking
   - Assessing coverage
   - Review pipeline and dashboard

3. Future developments

CJARS Pipeline begins with data intake

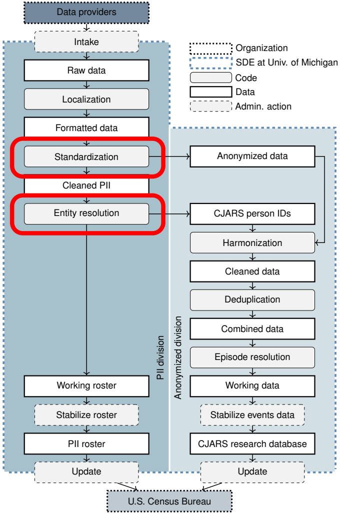
CJARS Criminal Justice Administrative Records System

- Metadata related to relational structure, data source, intake date, etc. are tied to a unique dataset identifier
  - `MI/St/DOC/20210422`
- The **Localization** stage brings all data into Stata format
- Most data processing is in Stata, with Python used for supplementary parsing and machine learning



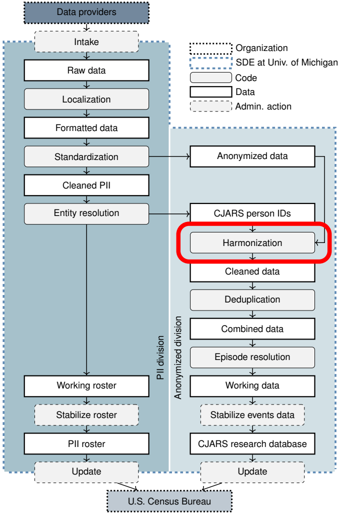
PII is extracted and linked

- **Standardization** includes race imputation
- **Entity resolution** algorithm matches PII within individual datasets, then reconciles matches at the state level, across all data sources (e.g. DOC, courts, county sheriffs)
    - See the working paper: `https://cjars.isr.umich.edu/entity-resolution-download/`
- PII is removed from source files and replaced with a **cjars_id**



Harmonization brings data into a national schema

- Each record in a CJARS table captures an individual criminal justice event
  - **Adjudication**: Individual charges, including conviction information
  - **Arrest**: Arrest and booking information, charge-level
  - **Incarceration, Probation, Parole**: Spells, including begin and end date
  - Documentation: `https://cjars.isr.umich.edu/data-documentation-download/`
- Tools for quality and consistency
  - TOC tool
  - State-level crosswalks



Four major data quality review processes

- Real-time review integrated into processing and code review
- Benchmarking against external data
- Assessing geographic and temporal coverage
- Review dashboard for aggregate statistics

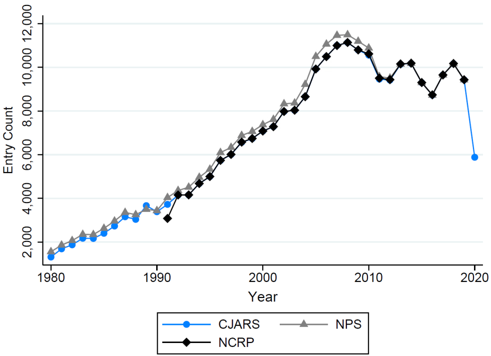
Managing data quality in production with real-time review

- Collaboration between data processing and data collection teams
- Integrated automatic quality checks
  - Data types, invalid dates
  - Reports generated by standard processing scripts
    - Percent of missing values
    - Distributions of dates and categorical variables

Benchmarking CJARS against public data

- Sources:
  - Annual Parole Survey and Annual Probation Survey (APS)
  - National Prisoner Statistics Program (NPS)
  - National Corrections Reporting Program (NCRP)
- High quality external data is available for incarceration, probation, and parole
- Adjudication and arrest data are harder to benchmark

Colorado - Raw count of incarceration entries by year

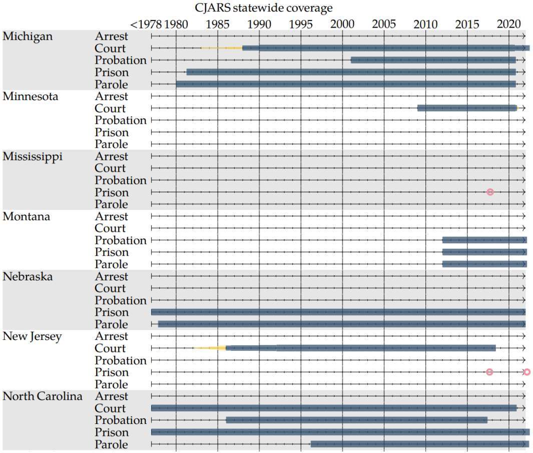


Assessing coverage

- We receive datasets with widely varying temporal and geographic coverage, from historical data to snapshots of current populations

- A data-driven process provides an initial guess at the coverage based on the distribution of key date variables

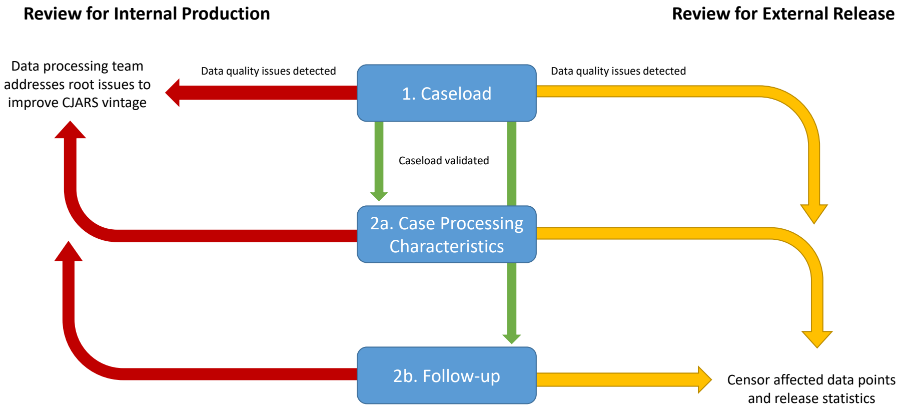
- The data collection team makes a full coverage assessment after harmonization



Review dashboard for aggregate statistics

- Three kinds of aggregate statistics:
  - **Caseloads** – per capita rates of incarceration, charges, convictions, etc.
  - **Case processing characteristics** – case processing time, average incarceration spell length, etc.
  - **Follow-up statistics** – recidivism; outcomes related to health and income (produced on the Census Bureau's IRE system)
- Two-stage pipeline:
  - Caseload statistics go through algorithmic and human review
  - Once a caseload statistic is validated, any other statistics that it supports are cleared for review
    - e.g. **Average incarceration spell length** for a jurisdiction and date range can only be reviewed once the respective **incarceration entries and exits** have been validated

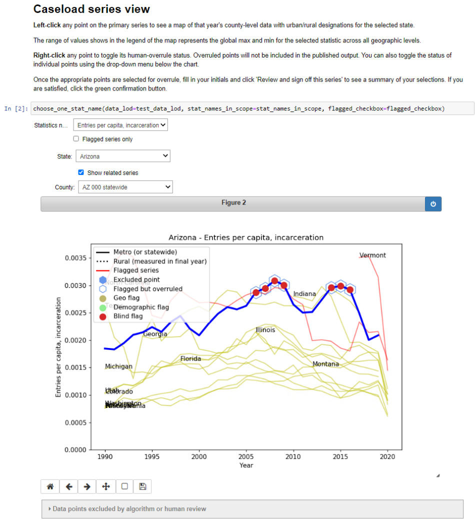
Dashboard supports internal review and external release



Review for Internal Production
Review for External Release
Data processing team addresses root issues to improve CJARS vintage
Data quality issues detected
Data quality issues detected
1. Caseload
Caseload validated
2a. Case Processing Characteristics
2b. Follow-up
Censor affected data points and release statistics

Automated review pipeline

1. Initial demographic review
2. An ensemble of heuristics and statistical checks flag individual data points
3. Human review of caseload statistics via interactive dashboard
4. Propagate results of caseload review to dependent statistics
5. Repeat the pipeline and human review process for case processing characteristics and follow-up statistics



Caseload series view

Left-click any point on the primary series to see a map of that year's county-level data with urban/rural designations for the selected state.

The range of values shows in the legend of the map represents the global max and min for the selected statistic across all geographic levels.

Right-click any point to toggle its human-overrule status. Overruled points will not be included in the published output. You can also toggle the status of individual points using the drop-down menu below the chart.

Once the appropriate points are selected for overrule, fill in your initials and click 'Review and sign off this series' to see a summary of your selections. If you are satisfied, click the green confirmation button.

```
In [2]: choose_one_stat_name(data_lod=test_data_lod, stat_names_in_scope=stat_names_in_scope, flagged_checkbox=flagged_checkbox)
```

Statistics n...: Entries per capita, incarceration

☐ Flagged series only

State: Arizona

☑ Show related series

County: AZ 000 statewide

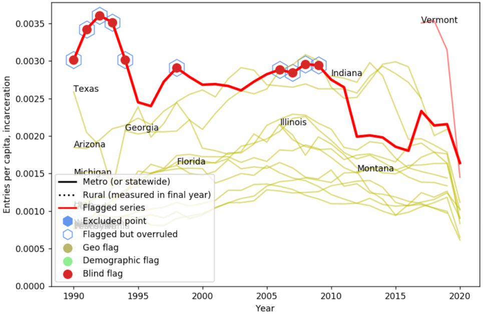
▸ Data points excluded by algorithm or human review

Flags in the automated review pipeline

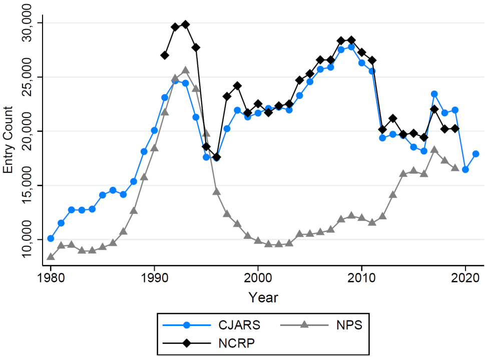
- Demographic check – look for jurisdictions with race, gender, and age profiles outside of reasonable upper and lower bounds
- 'Blind' check for outliers – look for extreme high and low values across all years and jurisdictions
- Autocorrelation, lag 1 – look for consistency of trend
- Urban/rural comparison – look for years in which the typical relationship between urban and rural counties changes
    - Aggregating counties into bins by size

Reviewing flagged series with multiple methods

North Carolina - per capita incarceration entries

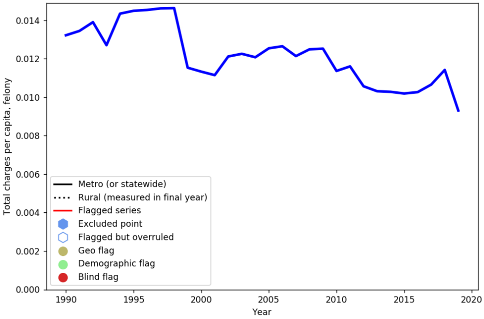


North Carolina - Raw count of prison entries

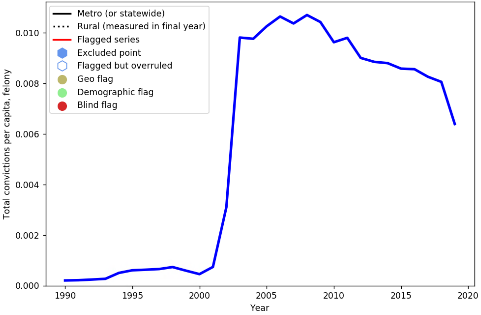


Identifying coverage issues within and across counties

Kent County, MI - Felony charges per capita

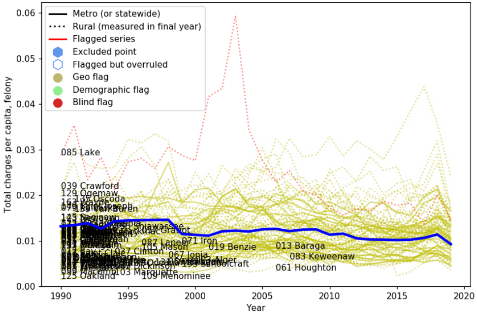


Kent County, MI - Felony convictions per capita

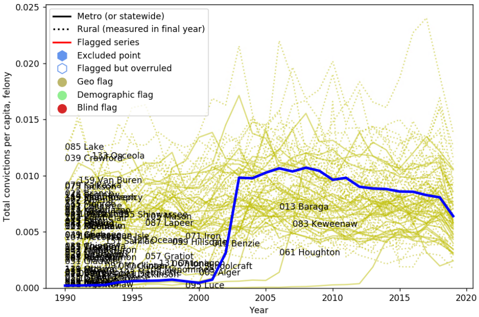


Identifying coverage issues within and across counties

Kent County, MI - Felony charges per capita

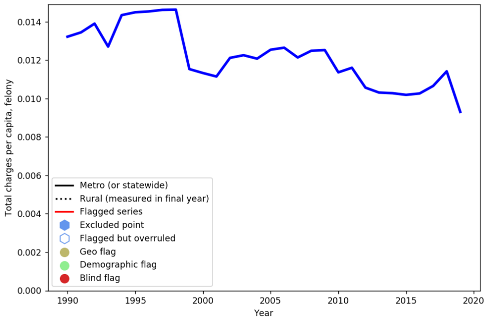


Kent County, MI - Felony convictions per capita

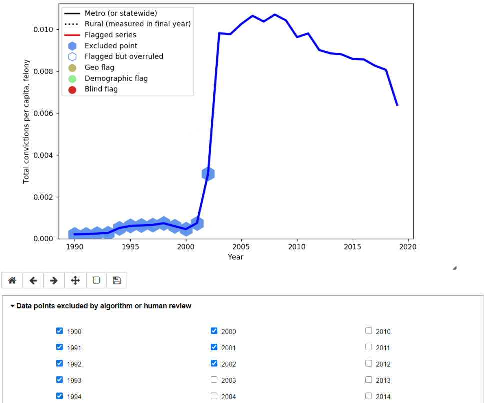


Censoring statistics for public release

Kent County, MI - Felony charges per capita



Kent County - Felony charges per capita, with dashboard interface



▾ Data points excluded by algorithm or human review

☑ 1990
☑ 1991
☑ 1992
☑ 1993
☑ 1994

☑ 2000
☑ 2001
☑ 2002
□ 2003
□ 2004

□ 2010
□ 2011
□ 2012
□ 2013
□ 2014

Future Developments

CJARS Criminal Justice Administrative Records System

- We are currently migrating to a new system designed to improve data processing and quality review in a variety of ways
  - Storing all metadata in JSON files, which facilitates:
    - Centralized schema for generating and validating variables
    - Browser-based interfaces for managing dataset metadata and data quality review checklists
  - Offloading space-intensive string variables into a SQL database
  - Managing all core data processing tasks with Python for improved parallelization, modularity, and logging
  - Improved tools for automatically parsing PII
- Learn more at `http://cjars.isr.umich.edu`