A Brief Guide on California's UCR Crime Data by City

Victor Kilanko

March 17, 2025

Introduction

- This project is based on *"A Note on the Use of County-Level UCR Data"* by Michael D. Maltz and Joseph Targonski (2002).
- Highlights potential data pitfalls and guides data manipulation efforts.
- Dives into CA city-level crime data, discusses potential issues, and provides a clean city-level crime data.

Key Issues with County-Level UCR Data

- County-level crime data have significant gaps, and the imputation methods used to fill these gaps are inadequate and inconsistent.
- City-level crime data do not suffer from the same imputation issues as county-level data.
- The county-level crime data were aggregated to the county level and archived at the National Archive of Criminal Justice Data (NACJD). The data were based on the "Crime by County" data file provided by the FBI's Uniform Crime Reporting (UCR) Program.
- Some law enforcement agencies (LEAs) did not report crime data, leading to FBI imputation methods that might render county-level crime data widely inaccurate.

Structure of the Paper (*"A Note on the Use of County-Level UCR Data"*)

- **Part 1:** The FBI's UCR Program.
- **Part 2:** Characteristics of the crime data.
- **Part 3:** Imputation procedures used by the FBI and NACJD.
- **Part 4:** How imputed data could lead to erroneous analysis (example of the MGLC analysis).

Part 1: The FBI's UCR Program

- The UCR Program started in 1930 to create a more accurate crime data representation.
- The UCR is voluntary. Crime reports were submitted by police agencies to the FBI, or submitted to a state agency (State UCR Program) which compiles and sends to the FBI.
- The FBI has no control over the reliability, accuracy, consistency, timeliness, or completeness of the data.
- Not all police agencies submit 12 months of crime data due to budget issues, national disasters, personnel changes, inadequate training, or no crime to report.
- Since the 1960s, the FBI has been filling in the omissions to make for their year-to-year trends.

Part 1: The FBI's UCR Program Cont'd

- The FBI imputes data at the agency level, but doesn't publish the imputed figures for agencies or counties. The figures are simply used to aggregate to the state, regional, or national level.
- They are not aggregated to the county level as they would be inaccurate because a single non-reporting agency may form a substantial part of a county's crime number.
- The FBI has a "Crime by County" file that includes crime data per reporting jurisdiction, months reported, and jurisdiction's population for agencies that reported.
- No population for any jurisdiction with no crime report. Hence, the county population may not be the actual total county population, just the population of all the reporting jurisdictions in the county.
- For cities in several counties, crime numbers per county are based on the population proportion of the city in each county.
- NACJD aggregated this FBI Crime by County file data to the county level which could be very misleading.

Part 2: Data Sources and Characteristics

- The authors (Maltz and Targonski, 2002) used original FBI-collected data from the National Consortium on Violence Research (NCOVR).
- Agency-level population data in the database was sourced from the Census Bureau, starting in 1980.
- The authors compared NACJD and NCOVR crime reports per county per year from 1980 to 1992.
- The crime counts were nearly identical, but discrepancies existed due to:
    - NACJD's allocation of crime counts from statewide agencies to each county based on county population proportions.
    - FBI dataset modifications over time, such as reporting deadlines.

Challenges in Population Data

- NACJD and NCOVR also had varying population data due to:
  - Double-counting of agency populations.
  - Different estimates of agency populations.
  - Agencies with overlapping jurisdictions.
- Double-counting occurs when a jurisdiction has two reporting agencies, and raw UCR data does not account for this. NACJD avoided this issue because FBI's Crime by County data accounted for it, while NCOVR used census data.
- Some agencies operate in multiple counties, leading to varying population estimates depending on the estimation method used by different agencies (NACJD vs. Census Bureau).
- Overlapping police jurisdictions exist in cases where cities have police department, military barracks police, and campus police. The FBI treats barracks and campus police as zero-population jurisdictions. If these zero-population jurisdictions fall within a city, the city's police department already reports its total population, but the crime figures of these agencies are added to the city's crime numbers.

Part 3: Imputing UCR Data - FBI's Method (1994 till Date)

- The FBI fills data gaps with estimates to allow year-to-year comparisons of state, regional, or national crime figures.
- Omissions are filled at the agency level.
- If an agency reports ≥ 3 months of data, the total crime for the year is calculated as:

$$C \times \frac{12}{N}$$

  where C is total crime for the reported period and N is the number of months reported.
- If an agency reports ≤ 2 months of data, crime counts are imputed using:

$$C_s \times \frac{P_a}{P_s}$$

  where C_s is crime count of a similar agency, P_s is similar agency's population, and P_a is the population of the actual agency (all for the year in question).

Part 3: Imputing UCR Data - FBI's Method (1994 till Date) Cont'd

- Every (reporting) jurisdiction and its population are included in aggregate statistics.
- No imputation is done for zero-population or statewide agencies with missing monthly crime data.
- According to Kaplan (2024), this remains the current imputation practice.

Part 3: Imputing UCR Data - NACJD Method (1977-1993)

- NACJD aimed to provide a cross-sectional view of crime rates across the U.S.
- NACJD aggregated jurisdiction-level data to the county level.
- Agencies with < 6 months of reported data were excluded (both crime counts and population).
- Agencies with $6 - 11$ months of reported data had values adjusted using:

$$C \times \frac{12}{N}$$

- Statewide agencies' crime was allocated to counties based on their share of the state population.
  - If state crime data were mostly from rural counties (with lesser population) but allocated proportionally across all counties, rural areas faced undercounting, while urban areas had overcounts.
- NACJD did not impute data for zero-population agencies.

Part 4: The Problem with Imputation Methods in MGLC

- The first edition of MGLC (covering 1977-1992) used NACJD's imputation method but relied on census population data.
- MGLC assumed that missing agencies dropped by NACJD (for under-reporting) had no crimes.
- As a result, the aggregate county-level crime rates in MGLC were substantially lower than actual rates.
- Crime rates were still calculated based on the entire county census population, which introduced inaccuracies.
- The second edition of MGLC (covering 1993-1996) did not account for the imputation method change after 1994, but this paper focuses on the first error.

Illustration of Imputation Errors in MGLC (pg. 312)

- Example: County A's crime rate calculation for 1989, 1990, and 1991.
- MGLC used Census Bureau data for county population (5000), while NACJD dropped agencies with < 6 months of reports.
- County A has agencies 1, 2, 3, and 4. Assume agencies 3 and 4 are zero-population agencies.
- 1989:
  - If all agencies reported a total of 120 crimes according to NACJD.
  - Crime rate per 100,000: $\frac{120}{5000} \times 100000 = 2400$
- 1990:
  - Agencies 1 and 2 reported less than 6 months and were dropped.
  - Agencies 3 and 4 reported a total of 5 crimes.
  - Crime rate per 100,000: $\frac{5}{5000} \times 100000 = 100$
- If MGLC relied on NACJD for population data, the population would have been dropped as well, making the crime rate: $\frac{5}{0} \times 100000 = \textit{Indeterminate}$.

Illustration of the Imputation Errors Cont'd

- Assuming that the year was later than 1994, the crime data would be based on the FBI imputation method, all the population would still be used (5000), and if reported months were ≥ 3, the formula $C \times 12/N$ would be applied to estimate the total crime count for each jurisdiction. This means that the same crime reports that were dropped in 1990 would have been used in 1994 and beyond (inconsistent).
- 1991:
  - Agency 1 reported 60 total crimes for all months, with a population of 3000 (from Census Bureau).
  - Agency 2 reported less than 6 months and was dropped (NACJD rules).
  - Agencies 3 and 4 reported a total of 10 crimes.
  - Crime rate per 100,000 (NACJD): $\frac{70}{3000} \times 100000 = 2333$
  - Crime rate per 100,000 (MGLC): $\frac{70}{5000} \times 100000 = 1400$
  - What if Agency 2 wasn't dropped?

Implications of Imputation Issues

- The dropping policy (NACJD's crime counts) creates a false perception of declining crime rates. FBI's imputation method in adjusting crime counts doesn't help either.
- These policies significantly alter crime per county. Hence, why UCR county-level data is misleading.
- Inconsistent imputations and dropping. Agency 1 can be dropped in 1991, added in 1992, and imputed differently in 1994.
- Now, imagine if over 50% of county-level data points in CT, IN, and MS are missing crime data from more than 30% of their population.
- Additionally, if 13 states have more than 20% of data points with missing crime data from over 30% of the population.
- The states with the greatest data gaps coincided with changes in their Right-to-Carry (RTC) laws.
- City-level data remain largely unaffected as long as missing data haven't been imputed.

About My Data- CA UCR Data

- The Department of Justice (DOJ) Criminal Justice Statistics Center (CJSC) collects crime and clearance data reported by LEAs across the state in a panel dataset.
- This data is submitted as part of the FBI UCR Program and includes both the number of actual offenses and the number of clearances, following UCR guidelines.
- There is no record of imputations of missing data by the DOJ CJSC at the LEA level.
- There is no population data.

UCR Crime Offenses

- The eight criminal offenses collected in the UCR Program were chosen due to their seriousness, frequency, geographic pervasiveness, and likelihood of being reported.
- These offenses include criminal homicide, rape, robbery, aggravated assault, burglary, larceny-theft, motor vehicle theft, and arson (ranked in hierarchy).
- These are 8 of the Part 1 offenses, excluding human trafficking (commercial sex acts or involuntary servitude).
- Official definitions of each crime can be found in the CJIS Division UCR Program User Manual (2013).
- Classification examples include:
    - Carjackings are categorized as robbery, not motor vehicle theft.
    - Robbery via housebreaking is classified as burglary.
    - Robbery without force or threat is larceny-theft.

Crime Data Reporting

- The number of reported homicide, rape/forcible rape, and aggravated assault crimes represents known victims.
- The number of robbery, burglary, larceny-theft, motor vehicle theft, and arson represents known incidents.
- One arson incident could result in multiple homicide victims.
- Hierarchy Rule: When multiple offenses occur in an incident, only the highest-ranking Part 1 offense is reported, except for justifiable homicide, motor vehicle theft, human trafficking, and arson, which are always counted. Hence, the crime data may be intrinsically undercounted.
- If an offender commits a series of offenses with a separation of time and place, each crime is recorded as a separate incident. Otherwise, it is treated as a single criminal transaction.

Crime Data Reporting and Transition to NIBRS

- In 2016, the FBI announced a transition to National Incident-Based Reporting System (NIBRS)-only by January 1, 2021.
- The California DOJ developed the California Incident-Based Reporting System (CIBRS) to meet new federal reporting standards.
- Since 2021, California's crime data has been a mix of summary and incident-based reporting (IBR) formats.
- A standardized method converts IBR data to summary data for comparison and trend analysis.
- Beginning in 2021, larceny theft value categories are based on standardized IBR to SRS data conversion which eliminates the $200 - $400 category.

California UCR Analysis (January 2017 - December 2023)

- Analyzing reported crime data across California from 2017 to 2023.
- Major concerns include overlapping jurisdictions, missing data, and negative values.
- Data cleaning involved handling local and county law enforcement overlaps, verifying jurisdictions, and addressing missing data.

Data Handling Protocol

- Overlapping jurisdictions can cause double counting (e.g., city police vs. county sheriff).
- Per the UCR Handbook (2004) and User Manual (2013):
  - City law enforcement agencies report crimes within city limits.
  - County/state LEAs report crimes occurring outside city limits.
- County sheriff reports are ignored when a city PD report exists.
- Excluded jurisdictions covering entire counties and where city values are unclear (e.g., Highway Patrol, BART, Parks).
- Zero-population jurisdictions' crime counts (e.g., universities, hospitals) are added to their respective cities after verification.
- For any missing data, the respective LEA is entirely dropped across the data. Hence, no data imputation.
- So, crime rate is based only on complete cities and their respective population.

Exact Order of Data Manipulation

- Created a pivot table (‘pivot_table_crime_data1‘) sorting data by county, agency, year, and month.
- Checked agencies per year/month for missing data or negative values in the violent and property crime sums.
- Some counties (e.g., Amador, Calaveras) lacked reports from certain cities.
- Any agency with missing data was entirely removed.
- Created a separate table (‘unique_county_agency1‘) to document jurisdiction-specific adjustments.
- Cleaned original data with the above documented adjustments:
    - Deleted county-level data.
    - Reassigned zero-population jurisdiction crimes to corresponding cities.
- Final check for negative values in all 8 major crimes:
    - Agencies with negative values were completely removed.
    - ‘deleted_agency.csv‘ contains removed agencies.
    - ‘final_cleaned_no_negatives.csv‘ contains the fully cleaned dataset.

Possible Downsides

- Human error in manually verifying jurisdictional coverage may exist.
- Incorrect classification (city-wide vs. county-wide reporting) could lead to misallocation of data.
- If such errors occur, corrections can be made by updating ‘unique_county_agency1‘ and re-running the data processing script.
- The possibility of dropping an entire city that is important to specific researchers.

Summary

- Previous research has highlighted the limitations of county-level crime data due to imputation issues.
- The California DOJ provides city-level or LEA-level data, but it still contains challenges such as overlapping jurisdictions, missing values, and negative values.
- This study manually examined LEA/jurisdiction data to resolve these issues, ensuring a cleaner dataset for crime researchers, policymakers, and econometricians.
- The analysis transparently presents the flaws in county-level data, the inconsistencies in CA DOJ LEA-level data, and the rigorous data handling protocol used to improve data reliability.