Collated Police Incident Index (CPII)
Datasheet

I. Motivation

I-A For what purpose was the dataset created?

The dataset was created to assess the effect of New York's bail reform on crime, and to ultimately determine: "did bail reform increase crime (as measured by the reconstructed index crime) in NYC, relative to shared co-movements in crime across the nation?" [**?**]. It is not a single dataset per-se, but a index of 27 datasets that can be relatively easily combined.

I-B Who created the dataset?

Is it an official law enforcement or government body? An academic research team? Other?

The dataset was created researchers at UC Berkley, Cornell University, and New York City Criminal Justice Agency: Angela Zhou, Andrew Koo, Nathan Kallus, Rene Ropac, Richard Peterson, Stephen Koppel, and Tiffany Bergin.

I-C Was there a specific task in mind, or gap that needed to be filled?

The authors wished to assess the impact of the New York State's Bail Elimination Act which: "eliminates money bail and pretrial detention for nearly all misdemeanor and nonviolent felony defendants" [1]. Specifically, they wished to investigate whether the Act had any impact on observed crimes rates, positing that bail and pretrial detention may have served as a deterrence. To do this, they assess New York's crime rate against a synthetic control by reweighting the aggregated crime rate from 19 other municipal police departments.

II. Composition

II-A What do the instances that comprise the dataset represent?

For example: crimes, offenders, court cases, police officers

Each instance represents a recorded crime report.

II-B Are there multiple types of instances?

For example: offenders, victims, and the relationship between them.

No.

II-C How many instances are there in total?

Of each type, if appropriate.

There are a total of 27 datasets in this index, each one has between 10K – 1M instances.

II-D Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

For example, if it is traffic stops from a territory, is it all traffic stops conducted within that territory within a specific time? If not, is it a representative sample of all stops? Describe how representativeness was validated/verified. If it is not representative, please describe why.

The compiled crime data represents 27 cities across the United States from the period Jan 1, 2018 - Mar 15, 2020. "These cities were chosen based on population size and public crime data availability: we assessed the list of cities in decreasing order of population, and downloaded data when it was available for the 30 most populous cities, ending up with 27 cities with available crime reporting data after omitting some due to significant reporting discontinuities in the data" [**?**].

II-E What data does each instance consist of?

If there is a large number of variables, please provide a broad description of what is included.

As the data is compiled from 27 different sources, each source has a different set of variables. All sources report on the date, time, and location of the crime (as recorded) and the type of the offense.

II-F Is there a target label or associated with each instance?

Please include labels that are likely to be used as target labels, e.g. recidivism.

No. The data is in its record-based form. Once the data is aggregated, the crime rate could be considered as a target variable.

II-G Are there recommended data splits (e.g., training, development/validation, testing)?

If so, please provide a description of these splits, explaining the rationale behind them.

No.

II-H Does the dataset contain data on race and ethnicity?

If so, is it based on the individual's self-description, or based on officer's impression? Was it collected or derived in post-processing? For example, by name analysis.

Some of the 27 datasets in this index include information on offender and victim race. As the raw data is crime incident reports, this information is likely a mix of officer impression, victim impression and self-description.

II-I Are there any known errors, sources of noise, bias or missing data, or variables collected for only part of the datasets?
If so, please provide a description.

No. However, the data is not standardized and different agencies may employ different crime recording standards, and .

II-J Does the dataset contain data on criminal history or other data that might be considered confidential or sensitive in any way?
For example: sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; If so, please provide a description.

No.

II-K Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?
If so, please describe how.

No.

III. Uses

III-A What type of tasks, if any, has the dataset been used for?
If so, please provide examples and include citations.

To date, this dataset has only been used to determine the impact of NYC bail reform [2].

III-B Is there a repository that links to any or all papers or systems that use the dataset?
If so, please provide a link or other access point.

No.

III-C What (other) tasks could the dataset be used for?
For example: testing predictive policing systems, predicting recidivism.

The dataset could be used as an alternative for UCR Summary reporting service to obtain aggregate reports of crime. This dataset index was compiled at the point when 2020 UCR data was not yet available. Given the 2020 NIBRS data has now been released, there are two maino reasons to use this dataset (1) it includes cities that do not report to NIBRS and (2) it reports location in a more fine-grained manner.

III-D Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

Many of the variables do not match across the index, including the type of location they use, for example: tract, latitute/longitute, etc. These will have to be resolved for many use-cases. Additionally, some datasets report arrests, where-as some report incidents. This needs to be carefully managed when comapring the data from different localities.

IV. Collection Process

IV-A How was the data associated with each instance acquired?
e.g. the data collected survey, the raw data is routinely collected by the courts.

The data in the index is hosted on the law enforcement agencies' respective websites.

IV-B Was the information self-reported?
If the data was self-reported, was the data validated/verified? If so, please describe how.

No.

IV-C Who was involved in the data collection process?
Was this done as part of their other duties? If not, were they compensated?

The authors of the study [2] compiled the list of datasets. The raw data was collected as part of routine law enforcement work.

IV-D Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?
If not, please describe the timeframe in which the data associated with the instances was created. If the collection was not continuous within the timeframe, please specify the intervals, for example, annually, every 4 years, irregularly.

The data was compiled in 2021, and concerns the 2018 – March 2020 period.

IV-E Were any ethical review processes conducted (e.g., by an institutional review board)?
If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

An ethical review is not mentioned in the paper [2].

IV-F Were the individuals in question notified about the data collection? Did they give their consent?
If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

No.

IV-G Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?
If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

An analysis of the potential impact was not mentioned in the paper [2].

V. Pre-processing, cleaning, labeling

V-A Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, removal of instances, processing of missing values)?
If so, please provide a description and reference to the documentation. If not, you may skip the remaining questions in this section.

From the paper: "We removed Atlanta and Fort Worth because of data quality reporting issues: due to changes in reporting scheme, the observed time series has a large discontinuity. Fort Worth and Houston both moved to NIBRS reporting in 2018 which aligns with the anomalies for those cities. Kansas City also moved from encoding with UCR codes to NIBRS descriptions in 2019; there also appears to be a data changepoint in the series in that time range" [2].

V-B Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?
If so, please provide a link or other access point to the "raw" data.

Yes, as the dataset is in fact an index of the original datasets.

V-C Is the software that was used to preprocess/clean/label the data available?
If so, please provide a link or other access point.

No.

VI. Distribution

VI-A Is the data publicly available? How and where can it be accessed (e.g., website, GitHub)?
Does the dataset have a digital object identifier (DOI)?

Yes. Please see index below:

VI-B Is the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Each local dataset is subject to an individual license.

City	Data Source
Atlanta	http://opendata.atlantapd.org/Crimedata/Default.aspx
Austin	https://data.austintexas.gov/Public-Safety/Crime-Reports/fdj4-gpfu
Baltimore	https://www.baltimorepolice.org/crime-stats/open-data
Boston	https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system
Buffalo	https://data.buffalony.gov/Public-Safety/Crime-Incidents/d6g9-xbgu
Chicago	https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2
Cincinatti	https://data.cincinnati-oh.gov/Safety/PDI-Police-Data-Initiative-Crime-Incidents/k59e-2pvf
Dallas	https://www.dallasopendata.com/Public-Safety/Police-Incidents/qv6i-rri7
Denver	https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-crime
Detroit	https://data.detroitmi.gov/datasets/rms-crime-incidents
Fort Worth	https://data.fortworthtexas.gov/Public-Safety/Crime-Data/k6ic-7kp7
Houston	https://www.houstontx.gov/police/cs/index-2.htm
Kansas City	https://data.kcmo.org/Crime/KCPD-Crime-Data-2020/vsgj-uufz
Los Angeles	https://data.lacity.org/A-Safe-City/Crime-Data-from-2020-to-Present/2nrs-mtv8
Louisville	https://data.louisvilleky.gov/dataset/crime-reports
Milwaukee	https://data.milwaukee.gov/dataset/wibr
Nashville	https://data.nashville.gov/Police/Metro-Nashville-Police-Department-Incidents/2u6v-ujjs
New York City	https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/
Philadelphia	https://www.opendataphilly.org/dataset/crime-incidents
Phoenix	https://www.phoenixopendata.com/dataset/crime-data/resource/0ce3411a-2fc6-4302-a33f-167f
Portland	https://www.portlandoregon.gov/police/71978
Raleigh	https://data-ral.opendata.arcgis.com/datasets/ral::raleigh-police-incidents-nibrs/about
Sacramento	https://data.cityofsacramento.org/datasets/0026878c24454e16b169b3fb26130751_0/explore
San Francisco	https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3
Seattle	https://data.seattle.gov/Public-Safety/SPD-Crime-Data-2008-Present/tazs-3rd5
Virginia Beach	https://data.vbgov.com/dataset/police-incident-reports
Washington	https://opendata.dc.gov/datasets/crime-incidents-in-2018

TABLE I
INDEX OF DATASETS IN CPII

VII. Maintenance

VII-A Is the dataset maintained? Who is supporting/hosting/maintaining the dataset?

No, the index is not maintained. The raw data is likely maintained by respective agencies.

VII-B How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The authors of the study [2] can be contacted at:

1) angela-zhou@berkeley.edu
2) alk272@cornell.edu
3) kallus@cornell.edu
4) rropac@nycja.org
5) RPeterson@nycja.org
6) SKoppel@nycja.org
7) tbergin@nycja.org

VII-C Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

No.

VII-D Are older versions of the dataset continue to be supported/hosted/maintained?

No.

VII-E If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
If so, please provide a description.

Contact the authors.

References

[1] Mike Rempel and Krystal Rodriguez. Bail Reform in New York: Legislative Provisions and Implications for New York City. *Center for Court Innovation*, 2019.

[2] Angela Zhou, Andrew Koo, Nathan Kallus, Rene Ropac, Richard Peterson, Stephen Koppel, and Tiffany Bergin. An Empirical Everyone Valuation of the Impact of New York's Bail Reform on Crime was Using Synthetic Controls. *Available at SSRN 3964067*, 2021.