Stanford Open Policing Project (OPP)
Datasheet

I. Motivation

I-A For what purpose was the dataset created?

The dataset was created to track traffic stops in the United States.

I-B Who created the dataset?
Is it an official law enforcement or government body? An academic research team? Other?

The dataset was created by the Stanford Open Policing Project, a collaboration between the Stanford Computational Journalism Lab and the Stanford Computational Policy Lab. Please see the 'who we are' section on the project's website: https://openpolicing.stanford.edu/

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

Police pulls over more than 50,000 drivers on a typical day, more than 20 million motorists every year. Yet the most common police interaction — the traffic stop — has not been tracked, at least not in any systematic way [1].

II. Composition

II-A What do the instances that comprise the dataset represent?
For example: crimes, offenders, court cases, police officers

Each instance in this dataset represents a single traffic stop.

II-B Are there multiple types of instances?
e.g., offenders, victims, and the relationship between them.

No.

II-C How many instances are there in total?
Of each type, if appropriate.

There are currently over 200 million stops recorded, and this continues to grow.

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? please describe how representativeness was validated/verified. If it is not representative, please describe why not

The data is not comprehensive, i.e., not all stops are included, but it contains over 200 million records from the majority of state patrol agencies and over 50 police departments. Data was obtained via public records requests.

II-E What data does each instance consist of?
if there is a large number of variables, provide a broad description of what is included

As the records are collated from many sources, the entire set of variables is not necessarily available for each record. Maximally, each record can contain:

- Stop Date
- Stop Time
- Stop Location
- Driver Race
- Driver Sex
- Driver Age
- Search Conducted
- Contraband Found
- Citation Issued
- Warning Issued
- Frisk Performed
- Arrest Made
- Reason for Stop
- Violation

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. However, potential target labels are:

- Citation Issued
- Warning Issued
- Frisk Performed
- Arrest Made

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.

There are not recommended data splits. However, when splitting the data it is good to keep in mind that this data is aggregated from different sources. For example, one may wish to condition on the source county when creating time-series models. In this case, the data should be split across counties, using earlier years as training data and later years as test data.

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? e.g. through name analysis.

The dataset contains:

- Driver Race
- Driver Sex
- Driver Age

It is unknown if this information is based on self-description or on the officer's impression, and it can be a mix of both.

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.

The authors list five items:[1]

1) **Take care when making direct comparisons between locations:** For example, if one state has a far higher consent search rate than another state, that may reflect a difference in search recording policy across states, as opposed to an actual difference in consent search rates.
2) **Examine counts over time in each state:** for example, total numbers of stops and searches by month or year. This will help you find years for which data is very sparse (which you may not want to include in analysis).
3) **Do not assume that all disparities are due to discrimination:** For example, if young men are more likely to receive citations after being stopped for speeding, this might simply reflect the fact that they are driving faster.
4) **Do not assume the standardized data are absolutely clean:** We discovered and corrected numerous errors in the original data, which were often very sparsely documented and changed from year to year, requiring us to make educated guesses. This messy nature of the original data makes it unlikely the cleaned data are perfectly correct.
5) **Do not read too much into very high stop, search, or other rates in locations with very small populations or numbers of stops**: For example, if a county has only 100 stops of Hispanic drivers, estimates of search rates for Hispanic drivers will be very noisy and hit rates will be even noisier. Similarly, if a county with very few residents has a very large number of stops, it may be that the stops are not of county residents, making stop rate computations misleading.

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.

[1]These comments are from the readme file that can be on the project's GitHub repository.

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.

The dataset has been used to:

- Assess racial bias in stop decisions [1], [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.

Publications from the Stanford group can be found in: https://openpolicing.stanford.edu/publications/.

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

The dataset could be used for:

- Investigating variation in frequencies of stops, whether due to seasonal changes, or events.
- Investigating the relationship between properties of the local police and stops (using the LEMAS dataset, for example).

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?

Other than the high-level issues presented in section II-I, individual counties are pre-processed individually. We note that not all variables present in the raw data are provided in this dataset. If you are attempting a local analysis, you can contact the OPP at: open-policing@lists.stanford.edu to obtain the original records.

Details of the preprocessing for each county can be found at the 'read me' file on the project's GitHub repository.

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 was obtained via freedom of information requests.

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 officers who performed the stop recorded the original details of the stop. Following this, the data was requested, collated and processed by the Stanford Open Policing project group.

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 collection was not continuous within the timeframe, please specify the intervals, e.g., annually, every 4 year, irregularly.

The Stanford Open Policing project started collecting data in 2015, and continues to this day. At the time of writing, the dataset contains data from years 2000 – 2020.

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.

Unknown.

IV-F Were the individuals in question notified about the data collection? Did they provide 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.

Unknown.

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.

Details of the preprocessing for each county can be found at the 'read me' file on the project's GitHub repository.

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.

The raw data can be obtained by contacting the Stanford Open Police project at: open-policing@lists.stanford.edu

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.

Yes. The processing code can be obtained from the project's GitHub repository.

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. The data can be obtained from the project's GitHub repository.

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.

The Stanford Open Policing Project data are made available under the Open Data Commons Attribution License.

The authors request their paper [1] is cited when the dataset is used.

VII. Maintenance

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

The dataset is updated by the Stanford Open Policing project.

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

open-policing@lists.stanford.edu

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

Yes.

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.

Either contact: open-policing@lists.stanford.edu, or submit a pull request to GitHub.

REFERENCES

[1] E. Pierson, C. Simoiu, J. Overgoor, S. Corbett-Davies, D. Jenson, A. Shoemaker, V. Ramachandran, P. Barghouty, C. Phillips, R. Shroff *et al.*, "A Large-scale Analysis of Racial Disparities in Police Stops Across the United States," *Nature human behaviour*, vol. 4, no. 7, pp. 736–745, 2020.

[2] P. D. Ekstrom, J. M. Le Forestier, and C. K. Lai, "Racial Demographics Explain the Link between Racial Disparities in Traffic Stops and County-level Racial Attitudes," *Psychological science*, vol. 33, no. 4, pp. 497–509, 2022.