Survey of Prison Inmates (SPI)
Datasheet

I. Motivation

I-A For what purpose was the dataset created?

The SPI was created to produce national estimates for the state and sentenced federal prison populations. It aims to provide a record and means of tracking inmate characteristics, such as demographics, background, and criminal history [1].

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 U.S. Bureau of Justice Statistics.

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

The dataset was created to be the first national periodic inmate survey in the United States, collecting detailed information pertinent to evolving issues within the criminal justice domain.

II. Composition

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

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

Each row corresponds to a survey response from a prison inmate.

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 24,848 inmates (20,064 state and 4,784 federal prisoners) in the 2016 SPI dataset.

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 dataset is a representative sample of the (over 18) U.S. prison population. Achieved by using a two-stage sample design, where state and federal prisons are selected in the first stage, followed by individuals selected from these facilities in the second stage. Samples are then weighted to account for non-response. Further details can be found in the SIP methodology white paper [1].

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.

Each instance broadly consists of:

- Current offense and sentence.
- Incident characteristics.
- Firearm possession and sources.
- Criminal history.
- Demographic and socioeconomic characteristics.
- Family background.
- Drug and alcohol use and treatment.
- Mental and physical health and treatment.
- Facility programs and rules violations.

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.

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.

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.

No.

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.

There are two known potential sources of error/noise: nonresponse (where the demographics of the respondents is significantly different to the non-respondents), and a coverage bias (where the sample population did not represent the target population). Non-response and post-stratification weights are provided to compensate for these.

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.

Yes, the dataset contains information on criminal history, sentencing, demographic and socioeconomic characteristics, family background, drug and alcohol use and treatment, and mental and physical health and treatment.

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.

Indirectly, by a comparing criminal history, demographic information, and sentencing information with other sources that are not de-identified.

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:

- Investigate the demographics and characteristics of inmates [2].
- Investigate specific inmate populations, including women, parents, and minorities [3], [4].
- Investigate the link between rural prisons and incarceration levels [5].
- Investigate the use and sources of firearms used in crimes [6].

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.

Yes. Please see here:
https://www.icpsr.umich.edu/web/NACJD/studies/37692

III-C What (other) tasks could the dataset be used for?

For example: testing predictive policing systems, predicting recidivism.

The dataset could be used to research counterfactual sentencing.

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?

No.

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 acquired via interview, as well as being linked to records maintained by other government agencies, such as criminal records.

IV-B Was the information self-reported?

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

Survey responses are self-reported. Data from official record are not.

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 data was collected by employees of the Bureau of Justice Statistics.

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.

SPI has released new data irregularly between 1974 – 2016. The latest release is from 2016.

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 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?

The individuals were notified: ”before the interview prisoners were informed verbally and in writing that their participation was voluntary and that all information provided would be held in confidence” [1].

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.

The only processing specified in the methodology is the non-response and coverage weighting [1].

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.

As the weighting is provided as a separate variable, the raw data is still accessible.

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.

N/A.

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:
https://www.icpsr.umich.edu/web/NACJD/studies/37692

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 license is not specified, but a citation and deposit requirement are listed:

Citation Requirement: Publications based on ICPSR data collections should acknowledge those sources by means of bibliographic citations. To ensure that such source attributions are captured for social science bibliographic utilities, citations must appear in footnotes or in the reference section of publications.

Deposit Requirement: To provide funding agencies with essential information about use of archival resources and to facilitate the exchange of information about ICPSR participants' research activities, users of ICPSR data are requested to send to ICPSR bibliographic citations for each completed manuscript or thesis abstract. Visit the ICPSR Web site for more information on submitting citations.

VII. Maintenance

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

Yes, by the Bureau of Justice Statistics.

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

The Bureau of Justice Statistics can be contacted at: askbjs@usdoj.gov

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?

Yes.

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.

No.

References

[1] L. E. Glaze, *Methodology: Survey of Prison Inmates, 2016*. US Department of Justice, Office of Justice Programs, Bureau of Justice ..., 2019.

[2] M. A. Peterson, H. B. Braiker, and S. M. Polich, *Who Commits Crimes: A Survey of Prison Inmates*. Oelgeschlager, Gunn & Hain Cambridge, MA, 1981.

[3] T. L. Snell, *Women in Prison: Survey of State Prison Inmates, 1991*. US Department of Justice, Office of Justice Programs, Bureau of Justice ..., 1994.

[4] L. M. Maruschak, J. Bronson, and M. Alper, "Parents in Prison and their Minor Children: Survey of Prison Inmates, 2016," *Publication no. NCJ-252645, US Department of Justice; Washington, DC*, 2021.

[5] S. R. Porter, J. L. Voorheis, W. Sabol *et al.*, "Correctional Facility and Inmate Locations: Urban and Rural Status Patterns," *Center for Administrative Records Research and Applications: US Census Bureau*, 2017.

[6] M. Alper and L. Glaze, *Source and Use of Firearms Involved in Crimes: Survey of Prison Inmates, 2016*. US Department of Justice, Office of Justice Programs, Bureau of Justice ..., 2019.