This README explains how to make best use of the data from the Stanford Open Policing Project. We provide an overview of the data, and a list of best practices for working with the data.

Our analysis code and further documentation are available at https://github.com/5harad/openpolicing.

Overview of the data file structure

For each state in the dataset, we provide data in three formats:

1. The data converted to our standardized format, with a single CSV for each state. We describe the meaning of each column below. It will be easiest to start with the standardized format for most analyses.
2. The raw data as originally received from the state. Raw data may be in a variety of formats — CSV, XLS, etc — and may contain multiple files.
3. The raw data converted to CSV format, with no other processing. There may be multiple CSV files.

The clean data are available for direct download from the Stanford Open Policing Project website; please contact us for access to the raw data.

Description of standardized data

Each row in the standardized data for each state provides information for one state patrol stop. All standardized data files contain the following columns. If a column cannot be computed using the data a state has provided, it is set to NA. Some states also have additional columns (e.g., an ID for the officer making the stop), which we do not use in our analysis, but which we include here because they might be useful to other researchers. These extra columns are explained in the state notes.

For several fields (e.g., driver_race) we include a “raw” column which records the original data values from which we infer standardized values. For example, driver_race_raw might be “White Hispanic” which we code as “Hispanic” in the standardized driver_race field. We include the raw columns because our data processing pipeline is extensive, requiring judgment calls and subjective decisions. We aim to make our data processing as transparent as possible. Other analysts may choose to process the raw data differently if their needs or judgments differ.

Column name	**Column meaning**	**Example value**
id	The unique ID we assign to each stop. Contains the state and year.	VT-2011-00012
state	The two-letter code for the state in which the stop occurred.	VT
stop_date	The date of the stop, in YYYY-MM-DD format. Some states do not provide the exact stop date: for example, they only provide the year or quarter in which the stop occurred. For these states, stop_date is set to the date	2011-11-27

	at the beginning of the period: for example, January 1 if only year is provided.	
stop_time	The 24-hour time of the stop, in HH:MM format.	20:15
location_raw	The original data value from which we compute the county (or comparably granular location) in which the stop occurred. Not in a standardized format across states.	Winooski
county_name	The standardized name of the county in which the stop occurred.	Chittenden County
county_fips	The standardized 5-digit FIPS code in which the stop occurred.	50007
district	In several states (e.g., Illinois) the stop county cannot be inferred, but a comparably granular location can. This comparably granular location is stored in the district column. Most states do not have this column.	ILLINOIS STATE POLICE 01
fine_grained_location	Any higher-resolution data about where the stop occurred: e.g., milepost or address. Not standardized across states.	90400 I 89 N; EXIT 15 MM90/40
police_department	The police department or agency that made the stop. Not in a standard format across states.	WILLISTON VSP
driver_gender	The driver's gender, as recorded by the trooper. M, F, or NA.	M
driver_age_raw	The original data value from which we compute the driver's age when they were stopped. May be age, birth year, or birth date. Not in a standard format across states.	1988
driver_age	The driver's age when they were stopped. Set to NA if less than 15 or greater than or equal to 100.	23
driver_race_raw	The original data value from which the driver's standardized race is computed. Not in a standard format across states.	African American
driver_race	The standardized driver race. Possible values are White, Black, Hispanic, Asian, Other, and NA, with NA denoting values which are unknown. Asian refers to Asian, Pacific Islander, and Indian. Native Americans/American Indians are included in the "other" category. Anyone with Hispanic ethnicity is classified as Hispanic, regardless of their recorded race.	Black
violation_raw	The violation committed by the driver, in the language of the original data. Not in a standard format across states.	Speeding (10–19 MPH Over Prima

	Some stops have multiple violations.	Facie Limit *)
violation	The violation committed by the driver, standardized into categories which are consistent across states.	Speeding
search_conducted	A TRUE/FALSE value indicating whether a search was performed.	TRUE
search_type_raw	The justification for the search, in the language of the original data. NA if no search was performed. Not in a standard format across states. Some states have multiple justifications for a search.	CONSENT SEARCH CONDUCTED
search_type	The normalized justification for the search. Where possible, this is standardized into categories which are consistent across states. For example, if something is clearly a consent search, search_type is referred to as “Consent”.	Consent
contraband_found	A TRUE/FALSE value indicating whether a search was performed and contraband was found. FALSE if no search was performed.	TRUE
stop_outcome	The outcome of the stop. Many states have idiosyncratic outcomes — for example, “CHP 215” in California — so this column is not standardized across states. “Citation” and “Warning” are the values which occur most commonly across states. If the stop has multiple outcomes, the most severe outcome is used. For example, if a stop resulted in a citation and a warning, stop_outcome would be “Citation”.	Citation
is_arrested	A TRUE/FALSE value indicating whether an arrest was made.	TRUE

Best practices

We provide some lessons we’ve learned from working with this rich, but complicated data.

1. Read over the state notes and the state processing code if you are going to focus on a particular state, so you’re aware of the judgment calls we made in processing the data. Taking a look at the original raw data is also wise (and may uncover additional fields of interest).
2. Start with the cleaned data from a single small state to get a feel for the data. Rhode Island, Vermont, and Connecticut are all load quickly.
3. Note that loading and analyzing every state simultaneously takes significant time and computing resources. One way to get around this is to compute aggregate statistics from each state. For example, you can compute search rates for each age, gender, and race group in each state, save those rates, and then quickly load them to compute national-level statistics broken down by age, race, and gender.

4. Take care when making direct comparisons between states. 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.
5. 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).
6. 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.
7. 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.
8. 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.

Our analysis only scratches the surface of what’s possible with these data. We’re excited to see what you come up with!

Below are notes on the data for each state. They are not intended to be a comprehensive description of all the data features in every state, since this would be prohibitively lengthy. Rather, they are brief observations we made while processing the data. We hope they will be useful to others. They are worth reading prior to performing detailed analysis of a state.

Arizona

Original format: Excel (xls)
Time period: 2011–2015 (there are some stops in 2009 and 2010, but stop counts are much lower).
Columns with no data:

- police_department
- driver_age

Data notes:

- Counties were mapped in two ways. First, we determined which counties the codes in the County field referred to by using the highways which appeared most frequently in each coded county. Second, for stops which had no data in the County field, we used the values in the Highway and Milepost fields to estimate where the stop took place. For this, we relied on highway marker maps (sources: https://azdot.gov/docs/business/state-milepost-map.pdf?sfvrsn=0 and http://adot.maps.arcgis.com/apps/Viewer/index.html?appid=4f19dfc238c44815b310bc72a9827bc2) to map the most frequently traversed highways, which covered the vast majority of stops. Using these two methods, we were able to map 95% of stops which had any location data (i.e., values in either County or Highway and Milepost), and 89% of stops overall.
- It would be possible to map the highway and mile marker data to geo coordinates, like we did in Washington.
- Data for violation reason is largely missing.
- VehicleSearchAuthority might provide search type but we lack a mapping for the codes. TypeOfSearch includes information on whom was searched (e.g., driver vs. passenger), but does not provide information on the type of search (e.g., probable cause vs. consent). ConsentSearchAccepted gives us information on search type for a small fraction of searches.
- There is a two-week period in October 2012 and a two-week period in November 2013 when no stops are recorded. Dates are sparse in 2009–2010.
- We also received a file with partial data on traffic stops pre-2009; this is not included in the cleaned dataset.
- Some contraband information is available and so we define a contraband_found column in case it is useful to other researchers. But the data is messy and there are multiple ways contraband_found might be defined, and so we do not include Arizona in our contraband analysis.

Extra fields:

- officer_id
- stop_duration
- road_number
- milepost
- consent_search
- vehicle_type
- ethnicity

California

Original format: csv

Time period: 2009-07 to 2016-06

Columns with no data:

- police_department
- driver_age
- stop_time
- fine_grained_location

Data notes:

- CHP districts roughly map to counties, so we mapped stops to counties using the map of CHP districts, which is included in the raw data. Some counties appear to have very high stop rates; this is because they have very small populations. It seems likely that the stops occurring in those counties are not actually the resident population.
- Driver age categories are included in the raw data; these cannot be mapped to granular values, so we cannot fill out the driver_age field.
- Very few consent searches are conducted relative to other states.
- Contraband found information is only available for a small subset of searches: the raw data can tell you if a probable cause search or a consent search yielded contraband, but cannot tell you if contraband was located during a search conducted incident to arrest. We therefore exclude California from our contraband analysis.
- Shift time is included, but is not sufficiently granular to yield reliable stop time.

Extra fields:

- ethnicity

Colorado

Original format: Excel (old data), |-delim (new data)

Time period: 2010 to 2016-03

Columns with no data: none

Data notes:

- The state did not provide us with mappings for every police department code to police department name.
- Arrest and citation data are unreliable from 2014 onward. Arrest rates drop essentially to zero.
- Counties were mapped using a dictionary provided by the agency. Denver County has many fewer stops than expected given the residential population; this is because it only contains a small section of highway which is policed by the state patrol.
- Rows represent violations, not stops, so we remove duplicates by grouping by the other fields.

Extra fields:

- officer_id
- officer_gender
- vehicle_type
- out_of_state

Connecticut

Original format: online, downloaded from https://data.ct.gov/view/baka-5j97

Time period: 2013-10 to 2015-03

Columns with no data: none

Data notes:

- Counties were mapped by running the cities in the Intervention Location Name field through Google's geocoder.
- Rows appear to represent violations, not individual stops, because a small proportion of rows (1%) report the same officer making multiple stops at the same location at the same time. We grouped the data to combine these duplicates. We don't want to be overly aggressive in grouping together stops, so we only group if the other fields are the same.
- While there is some search type data, a high fraction of searches are marked as “Other”, so we exclude Connecticut from our consent search analysis.
- While there is some violation data, we exclude Connecticut from the speeding analysis because it has too much missing data in the violation field.

Extra fields:

- officer_id
- stop_duration

Florida

Original format: Excel (xls)

Time period: 2010 to 2016-10. Format changes, but starts being stable in 2011.

Columns with no data:

- contraband_found
- police_department

Data notes:

- The raw data is very messy. Two different data sets were supplied, both with slightly different schemas. However, they were joined by uniquely identifying features. The second data dump goes until 2016, while the first only goes until 2015. The fields missing in the second data set are thus missing for some rows.
- There are many duplicates in the raw data, which we remove in two stages. First, we remove identical duplicate rows. Second, we group together rows which correspond to the same stop but to different violations or passengers.
- The original data has a few parsing errors, but they don't seem important as they are spurious new lines in the last 'Comments' field.
- The Florida PD clarified to us that both UCC Issued and DVER Issued in the EnforcementAction column indicated citations, and we consequently coded them as such.
- While there is some data on whether items were seized, it is not clear if these are generally seized as a result of a search, and we thus do not define a contraband_found column for consistency with other states.

Extra fields:

- officer_id
- officer_gender
- officer_age
- officer_race
- officer_rank
- out_of_state

Iowa

Original format: tab-separated (txt)

Time period: 2006 to 2016-04.

Columns with no data:

- driver_age
- search_conducted
- search_type
- contraband_found
- is_arrested

Data notes:

- The data separates warnings and citations. They are very different with respect to which fields they have available. Both contain duplicates. This happens when individuals receive more than one warning or citation within the same stop. We remove these by grouping by the remaining fields by the stop key and date.
- In some cases, there are multiple time stamps per unique (key, date) combination. In most of these cases, the timestamps differ by a few minutes, but all other fields (except for violation) are the same. In 0.1% of stops, the max span between timestamps is more than 60 minutes. In those cases it looks like the same officer stopped the same individual more than once in the same day.
- Only citations have 'Ethnicity', which only provides information on whether the driver is Hispanic. We therefore exclude Iowa from our main analysis because race data is lacking.
- Only (some) citations have county, the warnings only have trooper district. The mapping for the districts is provided in the resources folder. Counties were mapped by comparing the identifiers in the LOCKCOUNTY field with the cities in the LOCKCITY field.
- The codes in the county field represent counties ordered alphabetically.

Extra fields:

- officer_id
- out_of_state

Illinois

Original format: ~-separated (txt)

Time period: 2004 to 2015

Columns with no data:

- is_arrested

Data notes:

- The data is very messy. The presence and meaning of fields relating to search and contraband vary year by year. Caution should be used when inspecting search and hit rates over time. We exclude Illinois from our time trend marijuana analysis for this reason.
- For state patrol stops, there is mostly no information on the county of the stop. Instead, stops are mapped to districts (see the district column), which have a one-to-many relationship with counties. See the relevant map: http://www.isp.state.il.us/districts/districtfinder.cfm. There is one district (#15) with a lot of stops that does not directly map to counties, as it refers to stops made on the Chicago tollways. We use districts in our analysis.
- Counties for local stops were mapped by running the police departments in the AgencyName field through Google's geocoder.
- The search_type_raw field is occasionally "Consent search denied", when a search was conducted. This occurs because the search request might be denied but a search was conducted anyway. Many searches have missing search type data, so we exclude Illinois from our search type analysis.

Extra fields:

- stop_duration (only partially filled out)
- vehicle_type
- drugs_related_stop
- district

Massachusetts

Original format: csv

Time period: 2007–2016; there are some stops in 2005 and 2006, but numbers are extremely low.

Columns with no data:

- stop_time
- police_department

Data notes:

- The search and outcome fields are inconsistent. We take the most progressive interpretation: if one of SearchYN, SearchDescr or the outcome columns indicates that there was a search, we label them as such.
- While we define a contraband_found column in case it is useful to other researchers, it is sufficiently messy (there are multiple ways you might define contraband_found, and they are quite inconsistent) that we exclude it from our contraband analysis.
- Violation data is not very granular.
- Counties were mapped by running the cities in the CITY_TOWN_NAME field through Google's geocoder.

Extra fields:

- out_of_state

Maryland

Original format: Excel (xls)

Time period: 2007, 2009, 2011 to 2014-04. Except for 2013 and 2014, we do not have precise dates for stops.

Columns with no data:

- county_name, county_fips
- Different years have different amounts of data recorded, as noted below.

Data notes:

- The data is very messy. It comes from three different time periods: 2007, 2009-2012, 2013-2014. They all have different column and slightly different conventions of how things are recorded. We attempted to standardize the fields as much as possible.
- Time resolution of the data varies by year. Prior to 2013, data is reported annually. From 2013 onward, data is reported daily. So stop dates prior to 2013 are not precise to the nearest day and are just reported as Jan 1.
- Counties were mapped by running the police departments in the Agency field through Google's geocoder, but this does not work for state patrol stops, for which we have no county information.
- While there is information on violation, speeding stops constitute a very small fraction of stops compared to other states, and we therefore exclude Maryland from our speeding analysis.

Extra fields:

- out_of_state
- arrest_reason
- stop_duration
- search_duration

Michigan

Original format: csv

Time period: 2012 to 2016-01. The earliest stops date from 2001, but it only picks up consistently in 2012; prior til then stop numbers are much lower.

Columns with no data:

- driver_gender
- driver_age
- search_conducted
- search_type
- contraband_found

Data notes:

- The original data had some unquoted fields (VoidReason and Description) which had commas in them. We manually fixed these with a python script (scripts/convert_MI.py).
- Driver race data has more than 50% missing data, so we excluded Michigan from the analysis in the paper.
- The codes in the CountyCode field represent counties ordered alphabetically.
- Rows represent violations, not stops, so we remove duplicates by grouping by the other fields.

Extra fields:

- officer_id

Missouri

Original format: MDB

Time period: 2010 to 2015

Columns with no data:

- stop_time
- fine_grained_location
- county_name, county_fips
- driver_gender
- driver_age
- violation
- search_type
- stop_outcome
- is_arrested

Data notes:

- **The original data was aggregated**. There is detail on a number of fields (age, stop purpose, outcome) that is not usable as it is not cross-tabulated with the other fields.
- Because this is aggregate data, stop date is only precise to the nearest year, and is recorded as Jan 1 for all stops.
- Counties for local stops were mapped by running the cities in the city field through Google's geocoder, but there is no county information for state patrol stops.

Extra fields: none

Mississippi

Original format: Excel (xlsx)

Time period: 2013-01 to 2016-07.

Columns with no data:

- fine_grained_location
- stop_time
- search_conducted
- search_type
- contraband_found
- stop_outcome
- is_arrested

Data notes:

- Counties were mapped using the dictionary provided, which is added to the raw data folder. Counties are numbered alphabetically.
- There is no data on Hispanic drivers, so we exclude Mississippi from our main analysis.

Extra fields:

- officer_id

Montana

Original format: Excel

Time period: 2009 to 2016. Stop counts are much lower in 2009.

Columns with no data:

- police_department
- contraband_found

Data notes: none

Extra fields:

- lat
- lon
- ethnicity
- city
- out_of_state
- vehicle_year
- vehicle_make
- vehicle_model
- vehicle_style
- search_reason
- stop_outcome_raw

North Carolina
Original format: fixed-width
Time period: 2000–2015
Columns with no data: none
Data notes:

- Stop time is often unreliable — we have a large overdensity of 00:00 values, which we set to NA.
- The location of the stop is recorded in two different ways. Some stops have a county code, which can be mapped using the provided dictionary, which is included in the raw data. Other stops are only labeled with the state patrol district. Some districts map directly onto counties, in which case we label the stop with that county. However, some districts cover multiple counties. Stops in these districts can thus not be unambiguously mapped to a single county. In both cases, district of the stop is provided in the "district" column, providing granular location data for the vast majority of stops.
- Action is sometimes No Action or a similarly minor enforcement action even when DriverArrest or PassengerArrest is TRUE. In these cases, we set stop_outcome to be Arrest because the stop_outcome field represents the most severe outcome of the stop.
- search_conducted is TRUE if either the driver or passenger is searched. In 3.6% of cases, the passenger is searched. As their names suggest, driver_race, driver_gender, and driver_age always refer to the driver.

Extra fields:

- district
- search_basis
- officer_id
- drugs_related_stop
- ethnicity

North Dakota

Original format: Excel (xlsb)

Time period: 2010 to 2015-06.

Columns with no data:

- police_department
- search_type
- search_conducted
- contraband_found
- stop_outcome
- is_arrested

Data notes:

- The data contain records only for citations, not warnings, so we exclude North Dakota from our analysis.
- Rows represent individual citations, not stops, so we remove duplicates by grouping by the other fields.
- The stop_purpose field is populated by citation codes.

Extra fields:

- drugs_related_stop

Nebraska

Original format: MDB

Time period: 2002-2014.

Columns with no data:

- stop_time
- fine_grained_location
- county_name, county_fips
- driver_gender
- driver_age
- violation — it is included in the raw data, but not cross-tabulated
- search_type
- contraband_found
- stop_outcome — it is included in the raw data, but not cross-tabulated
- is_arrested — it is included in the raw data, but not cross-tabulated

Data notes:

- **The original data was aggregated**. It was grouped by stop reason, outcome and whether there was a search separately. Therefore, it is not possible to cross tabulate them together. We only use the last grouping.
- State and local stops are mixed together, but identifiable by the dept_lvl field.
- The data is by quarter, not by day. So all stop_dates are the first date of the quarter.
- For state patrol stops, there is a strange jump (Q1) and then dip (Q2–4) in the data for 2012. It looks like for 2012 all stops are recorded as happening in the first quarter.

Extra fields: none

New Hampshire
Original format: Excel (xls)
Time period: 2014-2015
Columns with no data:

- police_department
- search_type
- search_conducted
- contraband_found
- is_arrested

Data notes:

- The driver_race field was populated by hand-written codes that we manually decoded. They are prone to mislabeling and should be used with caution only. Also, a very high percentage of stops (>30%) are missing race data entirely. We map the most common codes, covering more than 99% of stops with data, but we do not interpret the long tail of misspellings because many of them are ambiguous, we do not want to make assumptions, and it does not significantly improve the data. We exclude this dataset from our analysis because it has too much missing race data.
- The stop_purpose field is populated by infraction codes. Code descriptions can be found here: http://www.gencourt.state.nh.us/rsa/html/.
- The driver_age field was not populated for the 2014.2 data set.
- Rows represent violations, not stops, so we remove duplicates by grouping by the other fields.

Extra fields:

- lat
- lon
- out_of_state
- aerial_enforcement

New Jersey

Original format: comma-separated

Time period: 2009 - 2016

Columns with no data:

- driver_age
- search_conducted
- search_type
- contraband_found
- is_arrested

Data notes:

- New Jersey data may be updated: we received the data very recently, and still have a number of questions we are waiting on the state to answer.
- New Jersey uses sofware produced by LawSoft Inc.: http://www.lawsoft-inc.com. There are two sets of data: CAD (computer aided dispatch, recorded at the time of stop) and RMS (record management system, recorded later). They have almost completely disjoint fields, and only RMS records have information on searches. We believe the data from the two systems should really be joined, but according to the NJSP there is not a programmatic way to do so. Therefore, we process just the CAD data, which appears to be the dataset which corresponds to traffic stops.
- In the CAD data, there are often multiple rows per incident. Some of these are identical duplicates, which we remove. For the remaining records, we group by CAD_INCIDENT, because the NJSP told us that each CAD_INCIDENT ID refers to one stop. We verified that more than 99.9% of CAD_INCIDENT IDs had unique location and time, implying that they did, in fact, correspond to distinct events.
- driver_race and driver_gender correspond to the race of the driver, not the passenger.
- Statutes are mapped using the traffic code: http://law.justia.com/codes/new-jersey/2013/title-39, where possible.
- The CAD records were mapped to a county by running the TOWNSHIP values through the Google geocoder.

Extra fields:

- officer_id
- out_of_state
- vehicle_make
- vehicle_model
- vehicle_color

Nevada

Original format: Excel (xls)

Time period: 2012-02 to 2016-05. Stop counts appear somewhat low in 2012, potentially indicating incomplete data.

Columns with no data:

- stop_time
- county_name, county_fips
- fine_grained_location
- police_department
- driver_gender
- search_type
- search_conducted
- contraband_found

Data notes:

- Nevada does not seem to record Ethnicity or have any records of Hispanic drivers, so we exclude it from our analysis.
- The violation field is populated by infraction codes.

Extra fields:

- drugs_related_stop

Ohio

Original format: comma-separated (txt)

Time period: 2010 to 2015

Columns with no data:

- police_department
- driver_age
- contraband_found
- stop_type

Data notes:

- The stop_purpose field is populated by infraction codes. The corresponding laws can be read here: http://codes.ohio.gov/orc/.
- There is no data for contraband being found, but a related field could potentially be reconstructed by looking at searches involving drugs and an arrest.
- Counties were mapped using the provided dictionary, which is included in the raw data folder.
- We cannot find disposition codes (in DISP_STRING) which clearly indicate whether a citation as opposed to a warning was given, although there is a disposition for warnings.
- The data contains stops of both type TS and TSA, standing for "traffic stop"" and "traffic stop additional". The latter have a higher search rate and tend to have additional information (i.e., ASINC_STRING is not NA). We include both types in analysis, as they do not appear to be duplicates (addresses and times do not match) and we do not have a clear reason to exclude either.
- While there is data on search types, they only include consent and K9 searches, suggesting a potential difference in recording policy (many other states have probable cause searches and incident to arrest searches, for example).

Extra fields:

- lat
- lon
- officer_id
- drugs_related_stop

Oregon

Original format: Excel (xlsx)

Time period: 2010 to 2016. Stop counts are much lower in 2015 and 2016.

Columns with no data:

- stop_time
- county_name, county_fips
- fine_grained_location
- police_department
- driver_gender
- driver_age
- violation
- search_conducted
- search_type
- contraband_found
- stop_outcome
- is_arrested

Data notes:

- There is basically no data, including no data on Hispanic drivers, so we exclude Oregon from our analysis.
- Counts for 2015 and 2016 are much lower than in earlier years.

Extra fields: none

Rhode Island
Original format: Excel (xls)
Time period: 2005–2015. Stop counts are considerably lower in 2005.
Columns with no data:

- county_name, county_fips
- fine_grained_location

Data notes:

- The stops are mapped to state patrol zones, but there does not seem to be a clear mapping to counties. We store state patrol zones in the "district" column and use this column in our granular location analyses.

Extra fields:

- stop_duration
- out_of_state
- drugs_related_stop
- district

South Carolina
Original format: |-separated (txt)
Time period: 2005 to 2016. Stop counts are low in 2016, suggesting potentially incomplete data.
Columns with no data:

- stop_time
- fine_grained_location
- search_type

Data notes:

- The police_department field is populated by state patrol agency.
- More data on local stops is available at http://afc5102.scdps.gov/SCDPS_Exweb/SCDPS/PublicContact/PublicContact-012. It is aggregated by race and age group — potentially scrapable if useful.
- While there is data on violation, many of the stops have missing data, so we exclude South Carolina from our speeding analysis.

Extra fields:

- lat
- lon
- officer_id
- officer_race
- officer_age
- highway_type
- road_number
- stop_purpose

South Dakota

Original format: Excel (xls)

Time period: 2012 to 2015-09. Stop counts are low in 2012, suggesting potentially incomplete data.

Columns with no data:

- police_department
- driver_age
- driver_race
- search_conducted
- search_type
- contraband_found
- is_arrested

Data notes:

- Race data is missing, so we exclude South Dakota from our analysis.
- Some county names were misrecorded and needed editing.

Extra fields:

- vehicle_type
- out_of_state
- drugs_related_stop

Tennessee

Original format: csv

Time period: 2006-03 to 2016-06. There is technically data going back as far as 1996, but the stop counts are extremely low.

Columns with no data:

- police_department
- fine_grained_location
- driver_age
- search_conducted
- search_type
- contraband_found
- is_arrested

Data notes:

- The data contain only citations, so we exclude Tennessee from our analysis.
- The codes in the CNTY_NBR field represent counties ordered alphabetically.
- It would be possible to map the highway and mile marker data to geo coordinates, as we did in Washington.

Extra fields:

- road_number
- milepost

Texas

Original format: MDB

Time period: 2006–2015

Columns with no data:

- police_department
- driver_age
- is_arrested

Data notes:

- There is evidence that minority drivers are labeled as white in the data. See: http://kxan.com/investigative-story/texas-troopers-ticketing-hispanics-motorists-as-white/. We remapped the driver race field as provided using the 2000 surnames dataset released by the U.S. Census. See the processing script or paper for details.
- We asked whether there was a field which provided arrest data, but received no clarification. There is data on incident to arrest searches, but this does not necessarily identify all arrests.
- Based on the provided data dictionary as well as clarification from DPS via email, we classify THP6 and TLE6 in HA_TICKET_TYPE as citations and HP3 as warnings.

Extra fields:

- officer_id
- lat
- lon
- driver_race_original

Virginia

Original format: custom format

Time period: 2006 to 2016-04.

Columns with no data:

- stop_time
- fine_grained_location
- police_department
- driver_gender
- driver_age
- violation
- search_type
- stop_outcome
- contraband_found
- is_arrested

Data notes:

- **The original data was aggregated**.
- The data is aggregated by week, not by day.
- Some rows have an absurdly high number of stops or searches. We have an outstanding inquiry on this, but for now it is assumed to be correct.
- Counties were mapped using the provided dictionary, which is included in the raw data folder.
- There are no written warnings in Virginia and verbal warnings are not recorded, so all records are citations or searches without further action taken. We, therefore, exclude Virginia from our analysis, because they do not record the same set of stops as other states.
- In the raw data, "Traffic arrests" refer to citations without a search. "Search arrests" refer to a citation and a search (either before or after the citation). "Search stops" refer to searches without a corresponding citation.

Extra fields:

- officer_id
- officer_race

Vermont

Original format: Excel (xlsx)

Time period: 2010-07 to 2016

Columns with no data: none

Data notes:

- Stop purpose information is not very granular — there are only five categories, and we have no way of identifying speeding.
- The search type field includes “Consent search — probable cause" and “Consent search — reasonable suspicion". It is not entirely clear what these mean; we cannot find analogues in other states.
- Counties were mapped by running the cities in the Stop City field through Google's geocoder.

Extra fields:

- officer_id

Washington

Original format: csv

Time period: 2009 to 2016-03

Columns with no data:

- police_department
- is_arrested

Data notes:

- Counties were mapped by doing a reverse look-up of the geo lat/long coordinate of the highway post that was recorded for the stop, then mapping that latitude and longitude to a county using a shapefile. Details are in the WA_map_locations.R script.
- We created an officer ID field based on officer name. Duplicates are possible if officers have the same first and last name, however this is unlikely.
- Arrests and citations are grouped together in stop outcome, so we cannot reliably identify arrests. There is data on incident to arrest searches, but this does not necessarily identify all arrests.

Extra fields:

- officer_id
- officer_gender
- officer_race
- highway_type
- road_number
- milepost
- violations
- lat
- lon
- contact_type
- enforcements
- drugs_related_stop

Wisconsin

Original format: |-separated (csv)

Time period: 2011 to 2016-05. There is some data in 2010, but stop counts are very low.

Columns with no data:

- driver_age

Data notes:

- The data come from two systems ("7.3" and "10.0") that succeeded each other. They have different field names and are differently coded. This is particularly relevant for the violation field, which has a different encoding between the two systems; in order to map violations, we used the dictionaries provided by the state for both systems.
- There are two copies of the data: warnings and citations. Citations seems to be a strict subset of warnings, with some citation codes being different.
- The police_department field is populated by highway patrol agencies. There are only 6 of them.
- There are very few consent searches relative to other states, suggesting a potential difference in recording policy.
- countyDMV field refers to the county of the stop, as the WI police clarified for us.

Extra fields:

- lat
- lon
- officer_id
- vehicle_type
- drugs_related_stop

Wyoming

Original format: Excel (xlsx)
Time period: 2011–2012
Columns with no data:

- search_type
- search_conducted
- contraband_found
- stop_outcome
- is_arrested

Data notes:

- Only citations are included in the data, so we exclude Wyoming from our analysis.
- The police_department field is populated by the state trooper division.
- The violation field is populated by violated statute codes.
- We found an external mapping of statute codes and provide them in the raw data.
- Some county names were misrecorded and required editing.
- Rows represent citations, not stops, so we remove duplicates by grouping by the other fields.
- contraband_found could potentially be derived from violation codes (drug / alcohol / weapons), but it would be less reliable and not necessarily comparable to how we defined contraband_found for other states.

Extra fields:

- officer_id
- drugs_related_stop