Cannabis convictions address data cleaning and geocoding

This document details the data cleaning and geocoding process used to map addresses associated with cannabis-related convictions in Connecticut from 1982-2020 to census tracts, in accordance with the parameters laid out in Public Act 21-1, An Act Concerning Responsible and Equitable Regulation of Adult-Use Cannabis. As of June 21, 2021, there were 288,661 convictions matching the statutory criteria.

Background

Public Act 21-1 defines a set of "disproportionately impacted areas" and targets its equity measures at those areas. The definition is as follows:

- "Disproportionately impacted area" means a United States census tract in the state that has, as determined by the Social Equity Council under section 22 of this act, (A) a historical conviction rate for drug-related offenses greater than one-tenth, or (B) an unemployment rate greater than ten percent;
- "Historical conviction count for drug-related offenses" means, for a given area, the number of convictions of residents of such area (A) for violations of sections 21a-267, 21a-277, 21a-278, 21a-279 and 21a-279a of the general statutes, and (B) who were arrested for such violations between January 1, 1982, and December 31, 2020, inclusive, where such arrest was recorded in databases maintained by the Department of Emergency Services and Public Protection;
- "Historical conviction rate for drug-related offenses" means, for a given area, the historical conviction count for drug-related offenses divided by the population of such area, as determined by the five-year estimates of the most recent American Community Survey conducted by the United States Census Bureau;[1]

The act requires that the Social Equity Council annually determine the list of disproportionately impacted areas:

- Not later than August 1, 2021, and annually thereafter, the council shall use the most recent five-year United States Census Bureau American Community Survey estimates or any successor data to determine one or more United States census tracts in the state that are a disproportionately impacted area and shall publish a list of such tracts on the council's Internet web site.[2]

Per statutory mandate, the definition of disproportionately impacted area applies to certain census tracts in the state based on state arrest and conviction data and unemployment data from the federal American Community Survey. Census tracts are small areas defined by the U.S. Census Bureau which allow for geographic tabulation of various statistics. Census tracts have an average of roughly 4,000 residents and generally approximate the concept of a neighborhood.[3] Connecticut has 833 census tracts as of the 2019 Census Bureau data.

To calculate the conviction rate, state arrest and conviction residential address data from the Department of Emergency Services and Public Protection was "geocoded" into latitude/longitude points on the map and then allocated to the census tracts in which the address lay. Geocoding is a process “for converting street addresses into spatial data that can be displayed as features on a map, usually by referencing address information from a street segment data layer.”[4]

To calculate the unemployment rate, data from the federal American Community Survey was used. The ACS is an annual survey that aims to sample 1 in 100 U.S. residents to provide fine-grained estimates of various statistical measures.[5] It includes a census tract-level estimate for the unemployment rate averaged over the previous 5 years.

---

[1] PA 21-1 § 1
[2] PA 21-1 § 22(i)
[3] https://www.census.gov/programs-surveys/geography/about/glossary.html#par_textimage_13
[4] https://support.esri.com/en/other-resources/gis-dictionary/term/ceeb3e0e-3276-4b0d-b660-b0c101aa704d
[5] https://www.census.gov/programs-surveys/acs/about.html

Process
The following steps were taken to geocode the convictions data:

1. **Data cleaning**. The addresses were examined for issues that could cause them to fail to be geocoded. The issues identified are summarized in the next section along with the data cleaning steps that were taken to address the issues where possible. A flag was added in the data to identify the issues that were documented (e.g. cases where there was no known address, the individual was unhoused, etc.).
2. **Geocoding**. The cleaned dataset was geocoded using a geocoder from Department of Emergency Services and Public Protection (DESPP), with address data used for 9-1-1 administration.[6] This process resulted in 263,107 records being geocoded. The DESPP geocoder was compared to other geocoding options and was determined to be the most accurate. For maximum accuracy, 7,421 records that were geocoded to a town not matching the listed town were discarded. As described below, This left 255,686 records successfully geocoded, or 89% of the underlying dataset.

Issues flagged with un-geocoded records	# records un-geocoded	% of un-geocoded records
No issue identified	13,982	42.40%
Geocoded into wrong town	7,421	22.50%
Non-CT state	6,174	18.72%
No known address	3,356	10.18%
Unhoused	1,045	3.17%
Hotel/Motel/Campground	296	0.90%
Correctional facility	287	0.87%
PO BOX	236	0.72%
Shelter	112	0.34%
Missing town and state	68	0.21%
Total	**32,977**	**100.0%**

3. **Aggregation**. All of the geocoded addresses were then allocated to a census tract using the 2019 TIGER/LINE shapefile from the U.S. Census Bureau that defines Connecticut census tracts.[7]
4. **Accounting for non-geocoded records.** Different jurisdictions have substantially disparate geocoding performance due to varying data quality. For example, for the city of Danbury, only 3,492 out of 4,044 records were geocoded successfully (86.4%). Waterbury, by contrast, had 17,182 out of 18,467 geocoded successfully (93.0%). Without accounting for non-geocoded records, tracts in towns with worse data quality would be effectively penalized. To match the intent of the legislation more closely, each non-geocoded record not located in a correctional facility was allocated proportionally across tracts in the town where it was located.
5. **Rate computation**. For each census tract with nonzero population, the conviction rate was calculated by dividing the final conviction count by the population, as defined in the 2019 American Community Survey. Census tracts with conviction rate greater than 0.1 or unemployment rate greater than 10% were marked as disproportionately impacted areas.

Data issues identified
The records that were not geocoded were analyzed and the following issues were identified:

[6] https://data.ct.gov/Public-Safety/Connecticut-9-1-1-Address-Points/m6xx-nb28
[7] A shapefile is a file containing definitions of geographic areas and associated data. https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2019.html

Issues that prevented geocoding of records: 4,705 records were identified that could not be geocoded because either: 1) no street address was given, 2) the individual was unhoused, 3), a PO Box was provided as the street address, or 4) the record was missing both town and state.

1. No known address
   - 3,356 records were flagged as having no known address.
   - These records included phrases such as “No known address”, “Unknown”, “No street address”, etc. in the street field.

2. Unhoused individuals
   - 1,045 records could not be geocoded because the individual was unhoused and no street address was provided.
   - These records included phrases such as “homeless”, “displaced”, “city streets”, etc. in the street field.

3. PO Box
   - 236 records listed a PO box in the street address field.

4. Missing town and state
   - 68 records had null values for the town and state field.
   - These addresses could not be geocoded and were flagged as “Missing town and state”.

Issues that could be addressed through data cleaning:

5. Apartment complex listed as street address
   - 1,767 records were flagged as being part of an apartment complex where the complex was listed in the street address field. For apartment complexes where at least 30 records were identified, these records were tagged with the name of the apartment complex and were assigned a placeholder address (below) to enable geocoding.

Apartment complex	# records identified	Placeholder street address	City
Father Panik Village	614	160 CHURCH ST	BRIDGEPORT
P. T. Barnum Apartments	348	451 BIRD ST	BRIDGEPORT
Marina Village	237	20 RIDGE AVE	BRIDGEPORT
Bellevue Square	225	1 MARY SHEPARD PL	HARTFORD
Monterey Village	106	133 MONTEREY PL	NORWALK
High Ridge Gardens	56	23 SCUPPO RD	DANBURY
Success Village	54	109 COURT D	BRIDGEPORT
Roodner Court	52	261 ELY AVE	NORWALK
Charles F. Greene Homes	45	98 HIGHLAND AVE	BRIDGEPORT
Meadow Gardens	30	49 MEADOW ST	NORWALK

6. Hotel/motel/campground listed as street address
   - 296 records were identified as referring to a hotel, motel, campground, etc.
   - These records included the name of the facility in the street address field, but did not include an actual street address.
   - These records were flagged as “Hotel/Motel/Campground”.

- To geocode these addresses, manual geocoding would be needed.

7. Shelter
   - 112 records were identified with addresses referring to a shelter. These records include either the word “shelter” or the name of the facility in the street address field, but did not include an actual street address.
   - These records were flagged as “Shelter”.
   - To geocode these addresses accurately, manual geocoding would be needed, and in many cases the addresses could not be identified.

8. Data cleaning on street address field
   - The following issues were identified and addressed in the street address field for the un-geocoded records:
     - Remove notes from street address field indicating “Last known address” or “LKA”.
     - Remove prefixes such as “X”, “X0”, “XW” from the street number.
     - Replace the word “half” with “1/2”.

9. Town name standardization
   - DESPP standardized town names as used in the Emergency-911 system database.

10. Misspellings
    - A manual review of the remaining uncategorized addresses indicated that many were misspelled, no longer existed, or appeared to have been recorded in the wrong town.

Issues that resulted in exclusion of records:

11. Non-CT state
    - 6,310 records were identified as being in a state other than CT. Some of these records were geocoded, but do not impact the determination of disproportionately impacted areas.

12. Correctional facility
    - 287 records were identified as referring to a correctional facility.
    - These records were flagged as “Correctional facility” and are not included in the final map in keeping with recent policy changes that indicate that people who are incarcerated should not be counted at the address of the correctional facility.

Datasets

The aggregated results are saved in a dataset on the Open Data Portal at this link.