QSS 20 Social Impact Practicum: Geocoding

Grant Anapolle, Sam Brant, Kate Christie, Eden Price

QSS 20. Final Class Status Update. 06.01.2021

Outline

- Motivation
- Research questions
- Data
- Methods
- Results thus far
- Next steps

Motivation

- Previous work has examined the relationship between year and number of violations.
- Previous work has also examined the relationship between frequency of violations and whether or not they were committed by a farm labor contractor.
- Interested in examining the other side: demographics of those who have had violations committed against them.

Research questions

- How does the percentage of Hispanic people in a tract correlate with the number of H2A investigations?
- How do the poverty rates in a tract correlate with the number of H2A investigations?
- How has the number of violations changed over the time period?

Data sources

- We used the Department of Labor Wage and Hour Division data for the H2A investigations and employer information
- We used Geocod.io API to get coordinates for the employer worksites
- We used demographics data from the American Community Survey for census-tract level variables
- Our units of analysis are H2A cases and the addresses associated with them. (When visualizing we may aggregate to the census tract level).
- Our primary fields of interest are cases, violations, addresses, and the tract demographic variables from ACS

Methods: Data Acquisition

- We obtained the Department of Labor data from https://enfxfr.dol.gov/data_catalog/WHD/whd_whisard_20210415.csv.zip
- We used an API to gather demographic data from the American Community Survey by census tract and the Geocod.io API to geocode the locations of the employers in the DOL data
- We also used the 'geopandas' and 'census' packages in Python for geocoding and reading in shape files containing tract boundaries

Methods: Data Cleaning

- We subset the DOL WHD data to H2A violations and the TRLA catchment states: TX, MS, LA, KY, AL, TN. We also subset to cases after 2015 to match on the census and demographics data.
- We use the DOL WHD columns case_id, street_addr_1_txt, st_cd, cty_nm, zip_cd, GEOID, h2a_violtn_cnt
- We use a geocoding function that creates columns for longitude and latitude for each employer and then we convert the dataframe into a GeoDataFrame
- We use the geopandas package to read in tract level coordinates and use a spatial intersection to match each employer to a census tract
- We use the census package to pull ACS demographics relevant to our research questions and contruct a GEOID using the state code, county code and tract code for each tract in the ACS data
- We merge the ACS demographics data and the employer violations data on GEOID

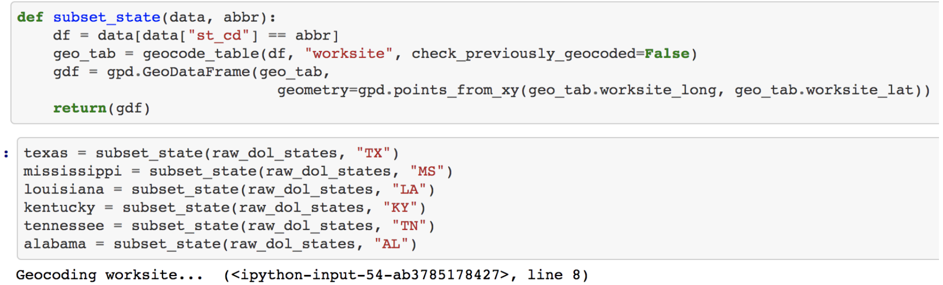
DOL Data Summary 1

	cases	addresses	violations	cmp_dollars	first_finding	last_finding
state						
KY	298	281	5058	1064416.70	2003-04-14	2020-12-15
LA	140	137	3658	1366161.03	2007-06-05	2020-03-11
MS	118	106	7301	1061084.56	2002-06-16	2020-09-02
TX	108	105	3395	2131764.65	2004-05-07	2020-10-15
TN	93	92	1978	652025.20	2001-03-14	2020-09-30
AL	49	45	440	137615.60	2002-05-01	2020-11-04

DOL Data Summary 2

	cases	addresses	violations	cmp_dollars
year				
2020	4	4	7	8859.40
2019	6	6	214	11786.30
2018	37	37	461	158420.58
2017	58	58	614	281048.66
2016	54	54	365	142213.60
2015	68	68	1948	342021.47
2014	63	63	1470	159240.23
2013	95	95	2659	762062.50
2012	74	72	1554	547837.50
2011	53	52	1655	611975.00
2010	80	80	1885	709100.00
2009	77	77	3085	2054357.50
2008	44	44	1036	183800.00
2007	26	25	1288	113830.00
2006	18	18	2829	265662.50
2005	18	17	232	11600.00
2004	12	12	216	17752.50
2003	13	13	126	12100.00
2002	5	5	157	14650.00
2001	1	1	29	4750.00

Geocoding employer addresses

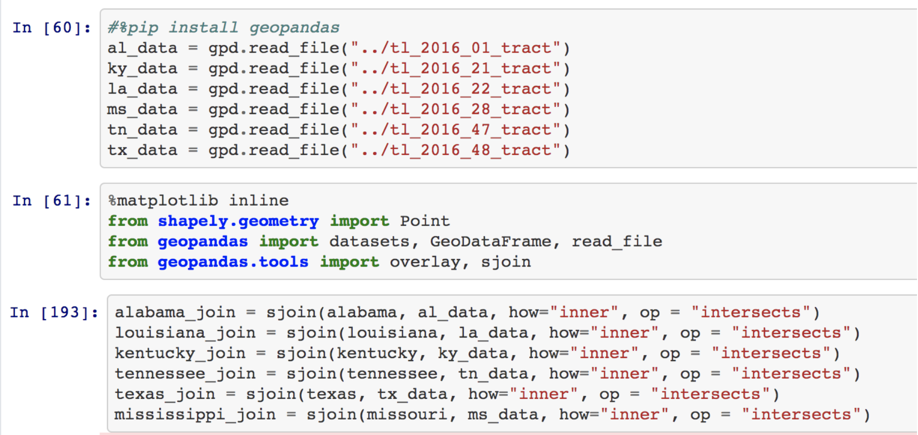


```
def subset_state(data, abbr):
    df = data[data["st_cd"] == abbr]
    geo_tab = geocode_table(df, "worksite", check_previously_geocoded=False)
    gdf = gpd.GeoDataFrame(geo_tab,
                             geometry=gpd.points_from_xy(geo_tab.worksite_long, geo_tab.worksite_lat))
    return(gdf)
```

```
texas = subset_state(raw_dol_states, "TX")
mississippi = subset_state(raw_dol_states, "MS")
louisiana = subset_state(raw_dol_states, "LA")
kentucky = subset_state(raw_dol_states, "KY")
tennessee = subset_state(raw_dol_states, "TN")
alabama = subset_state(raw_dol_states, "AL")
```

```
Geocoding worksite...  (<ipython-input-54-ab3785178427>, line 8)
```

Reading in tract coordinates and spatial intersection

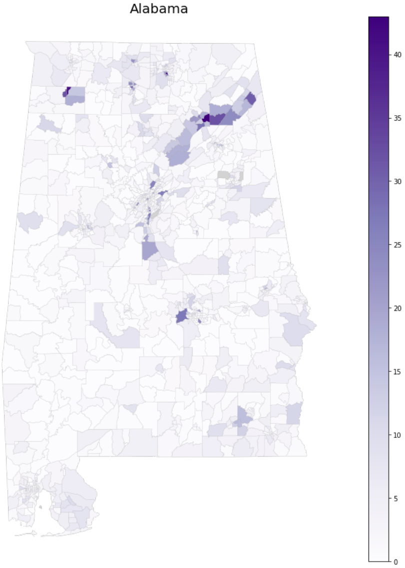


```
In [60]: #%pip install geopandas
         al_data = gpd.read_file("../tl_2016_01_tract")
         ky_data = gpd.read_file("../tl_2016_21_tract")
         la_data = gpd.read_file("../tl_2016_22_tract")
         ms_data = gpd.read_file("../tl_2016_28_tract")
         tn_data = gpd.read_file("../tl_2016_47_tract")
         tx_data = gpd.read_file("../tl_2016_48_tract")

In [61]: %matplotlib inline
         from shapely.geometry import Point
         from geopandas import datasets, GeoDataFrame, read_file
         from geopandas.tools import overlay, sjoin

In [193]: alabama_join = sjoin(alabama, al_data, how="inner", op = "intersects")
          louisiana_join = sjoin(louisiana, la_data, how="inner", op = "intersects")
          kentucky_join = sjoin(kentucky, ky_data, how="inner", op = "intersects")
          tennessee_join = sjoin(tennessee, tn_data, how="inner", op = "intersects")
          texas_join = sjoin(texas, tx_data, how="inner", op = "intersects")
          mississippi_join = sjoin(missouri, ms_data, how="inner", op = "intersects")
```

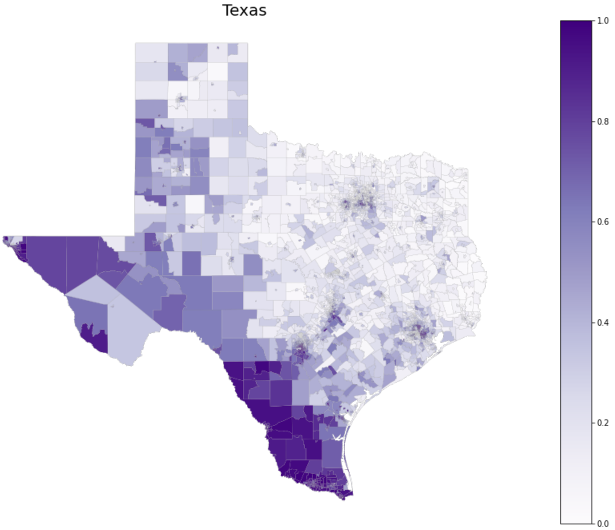
Methods: Visualization

- We are using matplotlib to visualize each of the 6 states in the TRLA catchment area
- We created heat maps for these states, the fill represents the percent Hispanic in each census tract and we are currently working on layering investigation locations on top of the heat maps in order to better understand the relationship between a Hispanic population and DOL investigations.

Results: Preliminary Visualizations



Results: Preliminary Visualizations



Next steps

- First, we will finish our preliminary visualizations by debugging our code to preserve the geometry column so that we can layer the demographics and the investigations on to one map
- Next, we will return to the ACS data to determine which columns are most useful in helping us answer our research questions and pull the remaining columns.
- Once we have all the columns and have converted them from raw numbers to percentages by tract, we can create the remaining visualizations.
- If time, we will also run bi-variate analyses to supplement our visualizations in order to more accurately answer our research questions
- In order to plot the violations over time, we will return to the census API to pull demographics across multiple years and create an animation or interactive visualization