QSS 20 Social Impact Practicum: Geocoding

Grant Anapolle
grant.j.anapolle.21@dartmouth.edu

Sam Brant
samantha.r.brant.23@dartmouth.edu

Kate Christie
katherine.e.christie.21@dartmouth.edu

Eden Price
eden.c.price.21@dartmouth.edu

Abstract

We are studying H-2A violations and the locations in which they tend to occur. We use data from the Department of Labor, which contains information about employers who had H-2A violations reported against them and the Census Bureau American Community Survey (ACS). We are primarily interested in the 6 TRLA catchment area states: Alabama, Kentucky, Louisiana, Mississippi, Texas, and Tennessee. The original DOL data includes violations from 2001 to 2020, we subset to violations in 2016 and later. We investigate our research questions by merging the DOL data with geographic coordinates from Geocod.io API and matching the coordinates to census tract shape files. We use ACS data for the 6 states and matched on the DOL data. Our final data frame contains demographic variables for all 10,870 census tracts in the 6 states, the employer addresses and coordinates for those with violations and the number of violations in each census tract. The primary research questions we are address in this paper are how does the Hispanic population correlate with H-2A investigations, how does the poverty rate correlate with H-2A investigations and how does the mobile or migrant population of a tract correlate with H-2A violations. We find moderate correlations across all three demographics but are unable to include census tracts with H-2A employers and no violations in the regression.

1 Introduction

In a typical year, the Department of Labor grants upwards of 250,000 H-2A visas to foreign nationals for temporary agricultural work in the United States. From the moment of recruitment, however, H-2A workers are highly vulnerable to abuses by their employers. Throughout QSS 20, our class has been preparing for the Social Impact Practicum in conjunction with Texas RioGrande Legal Aid (TRLA), a non-profit agency that provides legal aid to migrant workers. In line with TRLA's mission to protect vulnerable workers, we use data techniques to study H-2A violations in the six TRLA catchment states: Alabama, Kentucky, Louisiana, Mississippi, Texas, and Tennessee.

A high percentage of H-2A investigations by the DOL find violations against farm workers in the six states overseen by TRLA. Our project investigates the geographic and demographic information of found violations. In particular, we are interested in whether the locations of H-2A violations correlate with various key demographic variables. Using DOL and census data, we investigate how violations correlate with the size of the Hispanic population, the poverty rate, and the mobile or migrant population of a census tract. We hypothesize that the number of H-2A violations will correlate positively with all three demographic variables. Our geographic findings provide insight into the vulnerability of migrant workers on the H-2A visa program.

2 Related work

In March 2015, the US Government Accountability Office (GAO) published a report on the state of the H-2A and H-2B visa programs. The report investigated the number of workers on those visas, methods employers use to recruit those workers, and vulnerabilities and abuses that workers on those visa programs face. To investigate these issues, the GAO made use of numerous data sources. Data from the Department of State and Department of Homeland Security provided information

on H-2A and H-2B visas, including the number of visas given out, what occupations were filled through these visas, and demographic information about visa recipients. For information on abuses and vulnerabilities in the visa programs, the report used data from the Department of Labor, the Department of Justice, and the Department of Health and Human Services.

The GAO's investigation found that updated data practices would greatly benefit efforts to protect vulnerable visa workers. Although the DOL collects detailed information about debarred employers, that data was not used to screen new employer applications or shared with other enforcement agencies. This under-utilization of available data has resulted in many further preventable H-2A and H-2B violations.

In December 2020, the Economic Policy Institute (EPI) published a report on federal labor standards enforcement in agriculture. This report investigated the different investigations into federal labor standards violations against farmworkers in the United States. The EPI found violations in over 70 percent of investigations done by the Wage and Hour Division (WHD) of the Department of Labor (DOL).

The EPI noted that although the WHD is understaffed and underfunded, they were still able to find numerous violations of farmworkers rights. The present paper attempts to utilize the DOL data in a way that is useful in the work that TRLA does to protect H2-A workers and provide further insights into how to improve the H2-A program.

3 Data

Our primary data sources were the Department of Labor violations data and the American Community Survey demographics data. The Department of Labor Wage and Hour Division data provided information on employer address, number of violations and dates of investigations. This data was sourced from https://enfxfr.dol.gov/data_catalog/WHD/whd_whisard_20210415.csv.zip. The DOL data was subsetted to employers with at least 1 H-2A violation and our study focuses only on the following states: Alabama, Kentucky, Louisiana, Mississippi, Texas, and Tennessee. Geocod.io API provided the latitude and longitude of the employer addresses. We used the 'geopandas' package to load the 2016 TIGER/Line® Shapefiles of census tracts that were downloaded online from the census.gov website. The demographics data came from the Census Bureau American Community Survey (ACS), which was obtained using the 'census' package and an API.

Tables 1 and 2 show summary statistics of the DOL violations data by state and year. As Table 1 shows, there are more cases in the DOL data than there are distinct addresses because some addresses are repeated. There are two sources of repeat addresses. First, some employers have been investigated multiple times at the same address, as shown in Table 3. Second, different employers investigated separately may share an address, as shown in Table 4 if they were to rename their company to avoid further punishment, for example. As we proceeded with Geocoding H-2A violations, we kept in mind that addresses in the DOL data may repeat.

state	cases	addresses	violations	cmp_dollars	first load date	last load date
KY	78	76	1285	421809.20	2016-11-02	2021-04-15
MS	70	65	3182	598784.56	2016-11-07	2021-04-15
TX	58	58	1111	447777.15	2016-11-02	2021-04-15
LA	54	54	1389	320911.03	2016-11-02	2021-04-15
TN	40	39	863	238350.20	2016-11-02	2021-01-27
AL	29	27	230	83295.60	2017-07-18	2021-04-15

Table 1: Summary Statistics of Violations by State

4 Methods

We began by subsetting the DOL investigations data to the 6 TRLA states: Texas, Louisiana, Tennessee, Kentucky, Mississippi and Alabama, to employers that had at least 1 H-2A violation and to investigations that took place after Jan. 1, 2016. Our final data set contained 329 rows of employer addresses. In order to geocode the employer locations in the DOL dataset, we used the Geocod.io

year	cases	addresses	violations	cmp_dollars
2021	26	26	304	116322.28
2020	89	89	2527	775809.96
2019	44	44	1133	112069.55
2018	74	73	1430	569280.65
2017	63	62	1560	227482.80
2016	33	33	1106	309962.50

Table 2: Summary Statistics of Violations by Year in TRLA states

case_id	address	legal_name	load date
1708970	124 Jimmy Beckley Drive, Bruce, MS	Lewis M. Bailey, IV Farms, Inc.	2017-07-18
1775153	124 Jimmy Beckley Drive, Bruce, MS	Lewis M. Bailey, IV Farm Inc.	2018-03-24
1884517	124 Jimmy Beckley Drive, Bruce, MS	Lewis M. Bailey Farms, Inc.	2020-03-17

Table 3: Example of One Employer Investigated Multiple Times at the Same Address

case_id	address	legal_name	load date
1826254	9408 Mulligan Road, Owensboro, KY	Cecil Tobacco Company, LLC	2017-11-06
828480	9408 Mulligan Road, Owensboro, KY	Los Villatoros Harvesting LLC	2020-07-16

Table 4: Example of Different Employers Sharing an Address

API to match a latitude and longitude on to each employer address. Next, we were interested in determining which census tract each address was in.

We obtained census tract shape files from the census.gov website and used the 'geopandas' package in Python to load and read these files. Next, we used a spatial join to intersect the coordinates of the employer addresses on the census tract shape files to create a new column in the data set representing which census tract each employer address was located in and a GEOID which combines state code, county code and tract code for each row. We then spent time with the American Communities Survey data documentation to determine which columns would be useful in answering our research questions. In the final run of our code, we pulled the columns for the name of the tract, the number of people in each tract, and then three relevant columns on demographics. The first was "B01003_001E", which was the count of Hispanic individuals living in each census tract; this divided by the population in each census tract gave us the proportion of the census tract that was reported to be Hispanic. The second was "B06012_002E", the number of people at or below the poverty line in each tract. This number was also divided by the tract population to get a proportion. And finally, "B07007_005E", which is the population of foreign born, non-U.S. citizens who were mobile, or "movers" in the area in the past year. This column gave us the proportion of people who were assumed to be migrants or just mobile in the past year. After pulling the demographic by-tract data for each state, we concatenated the six data frames into one large dataframe containing all of the demographic statistics for each census tract in all 6 states. We created a GEOID for the ACS data by pasting together the state code, county code and census tract code to merge on.

Once we had a dataframe containing all of the employer locations, coordinates, census tracts, violation counts and GEOIDs and the ACS data containing all of the census tracts, GEOIDS and relevant demographics, we merged the two on the GEOID of each employer address. Our final dataframe contained demographic data for 10870 census tracts in 6 states. In order to preserve the demographic columns for tracts with no investigations, we replaced the missing values from the merge with 0. Next, we prepared our data for visualization.

We found the number of violations in each tract using the groupby function. Once we found the tracts that had at least one violation, we merged these rows back on to the geopandas data that contained the address coordinates in order to graph. To plot the data, we used the 'plotnine' package in pandas to do by-state visualizations. We created a function that allowed us to make visualizations for each state and on three of the variables we were interested in: the proportion of the Hispanic population in each census tract, the proportion of people at or below the poverty line, and then the proportion of foreign-born, non U.S. citizens who had migrated or were mobile in the past year.

State	Percent Hispanic R^2	Percent BPL R^2	Percent Mobile R^2
KY	0.852	0.738	0.879
AL	0.369	0.279	0.643
TN	0.277	0.563	0.379
TX	0.799	0.928	0.927
MS	0.706	0.517	0.789
LA	0.747	0.440	0.492

Table 5: The R-squared values obtained from the linear regression of H-2A violations and several demographics

In our visualizations, each census tract was shaded according to the proportion of the demographic variable and the number of violations were plotted, with size according to the amount of violations, as red circles on top of the map.

Finally, in order to statistically determine whether the correlations seen in our visualizations were significant, we used a linear regression on the number of H-2A violations and the proportion demographic for each tract. We obtained R^2 values for each state which indicated linearity of the fit (Table 5).

5 Results

We created 3 visualizations for each of the 6 states. Given that it was not feasible to pull all or many demographics from the ACS data, we chose three demographics that we believed might give some insight into the types of communities where violations are highly concentrated. The first maps the proportion of Hispanic people in the population and overlays points that represent the number of H-2A violations in each census tract. The second maps the proportion of people in each tract living at or below the poverty line and overlays points that represent the number of H-2A violations in each census tract. The final visualizations map the proportion of foreign born, non-U.S. citizens who were mobile in the past year and the points which represent the number of H-2A violations in each tract. For all of the maps, no dot indicates that there were no violations in the tract from 2016 - 2020.

Our descriptive statistics tables (Table 1) shows differences on the state level. As shown, Kentucky had the most H-2A investigations over the time period (78), and Alabama had the least (29).

Table 5 aggregates our findings from the linear regression. As shown, there is significant variation across the 3 demographics we investigated. When looking at the correlation between the proportion of Hispanic people in the tract and the number of H-2A violations, the correlation was moderately strong in Kentucky, Texas, Mississippi and Louisiana (.852, .799, .706, and .747) respectively. However, in both Alabama and Tennessee the correlation was weak (.369 and .277, respectively). There is also significant variation in the R^2 values for the correlation between the proportion living below the poverty line and the number of H-2A violations. We find that Texas has a very strong correlation of .928. Conversely, we find an R^2 value of .279 in Alabama. The other four states had moderate correlation, with R^2 values ranging from 0.44 and 0.73. Lastly, we found the smallest spread in coefficients for the proportion mobile demographic. Tennessee had the weakest correlation, with an R^2 value of .379, whereas Kentucky showed the strongest correlation, with an R^2 value of .879.

6 Discussion and Conclusion

While the scope of the present paper is severely limited, we offer a few relevant findings for the work of TRLA and provide opportunities for further study. We find moderate correlation across all three demographic variables and the H-2A violation count and show that Texas and Kentucky consistently have the strongest correlation coefficients. Our first research question investigates the relationship between Hispanic populations and H-2A investigations. We hypothesized that areas with higher Hispanic population would correlate with more H-2A violations given that the vast majority of H-2A visa holders are from South and Central America. Our results should not be interpreted as causal, but we find that in many states (Kentucky, Texas, Mississippi, and Louisiana) there was a moderate

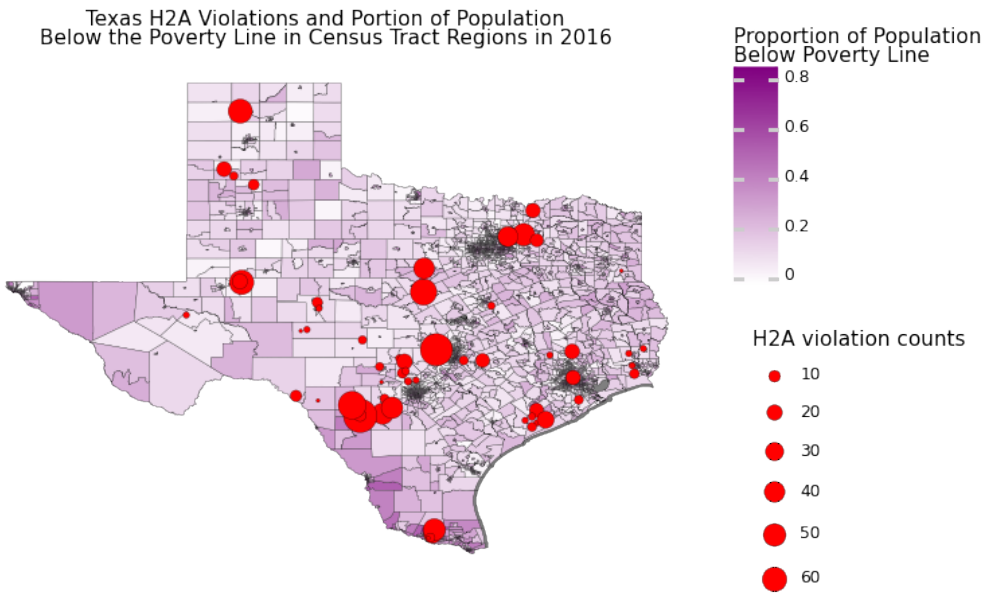

Figure 1: Map of Texas which is broken down by census tract. Shaded in are the proportions of each census tract that live at or below the poverty line, with the poorest areas indicated by darker purples. The demographics data is from the 2016 ACS survey. Plotted on top are the number of H2A violations, with larger circles indicating more violations at different worksites across Texas. This graph demonstrates one of the highest correlations, as the relationship between proportion of people living at or below the poverty line correlated to the number of H2A violations in that tract well, with an R^2 of .928

correlation between violations and the percent of the surrounding community that was Hispanic. In these four states, at least 70% of the data fit the linear regression model.

The next research question we were interested in investigating was relationship between the proportion of people at or below the poverty line and the violations. We hypothesized that employers in low-income areas would be more likely to violate the H-2A conditions and therefore have a higher amount of investigations. The mechanisms behind this assume that the H-2A employers are somewhat corrupt. We believe that workers in concentrated poverty would be less likely to report violations, in fear of losing their job, and that employers would be more likely to violate the H-2A conditions knowing that their workers would be less likely to report them. We find that only two states have a moderately strong correlation between poverty and violations: Kentucky and Texas (.738 and .928, respectively).

Our last research question addressed the relationship between the proportion of foreign-born, non-U.S. citizens who were mobile in the past year and violations. We were interested in this demographic because H-2A visas are given to migrant workers who may be reflected in this data. We hypothesized that higher concentrations of migrant workers would be strongly correlated with more violations. We find similar results to our 2 previous regressions. Some states had low $R\hat{2}$ values, including Louisiana and Tennessee while Texas and Kentucky showed strong correlation. Across the three variables, we found the strongest relationship between the mobility demographic and the violations, with an average R^2 value of .693, although there was not much variability. The first regression had an average $R\hat{2}$ of .625 across the 6 states and the second regression showed an average $R\hat{2}$ of .577 across the 6 states. Consistently, Texas and Kentucky showed the strongest correlation across all three variables.

When interpreting the findings of the regression, we must address a serious limitation in the calculations. Given the structure of the data frame, we were only able to include tracts with 1 or more

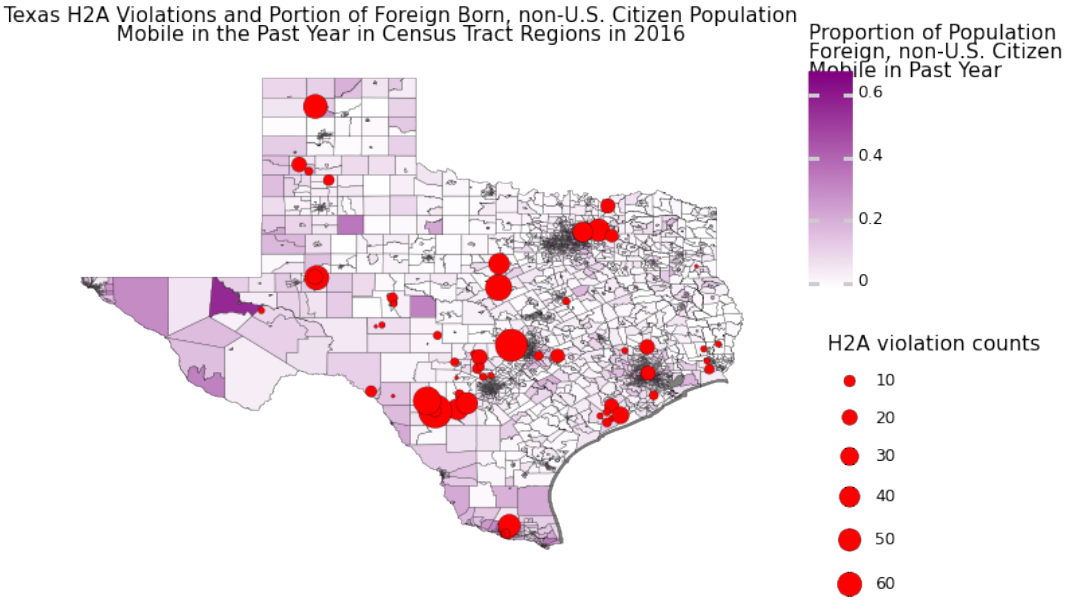

Figure 2: This is another map of Texas which is broken down by census tract. Shaded in are the proportions of each census tract that are foreign-born, non U.S. citizens who have been mobile or moved in the past year, with the areas with the most mobility indicated by darker purples. The demographics data is from the 2016 ACS survey. Plotted on top are the number of H2A violations, with larger circles indicating more violations at different worksites across Texas. This graph demonstrates one of the highest correlations, as the relationship between proportion of foreign-born, non-U.S. citizens who were mobile in the past year correlated to the number of H2A violations in that tract well, with an R^2 of .927

violation in our regression, excluding the calculation of all tracts with zero violations. We acknowledge that this limitation likely skewed our coefficients and that the inclusion of the zero-violation census tracts would likely produce significantly weaker correlations. For example, when looking at Figure 1 and 2, it is clear that there exist many census tracts with high proportions of Hispanic people or people living below the poverty line that had no H-2A violations. Tracts like these were not included in the regression and would likely have produced a weaker correlation coefficient. To reiterate, these were excluded from the analysis because the information on which tracts included H-2A employers with no violations was not accessible to us.

Beyond the limitations of the regression, this paper has other limitations worth mentioning. Firstly, the time frame of the project is a bit limited. We used census tract shape files from 2016 and 2016 ACS data. The investigations included were from 2016-2020, despite the original data containing employers investigated from 2001 onward. With more time, we would use ACS demographics from 2001 to 2020 to match with the violations from the corresponding years given that the three demographics we studied have likely been variable over the past 20 years. Next, our visualizations and analysis did not account for the situations in which one employer was subject to multiple investigations. While all of the investigations were counted in our final aggregation of by-tract investigations, we acknowledge that there is a substantive difference between employers with multiple investigations and employers with one. Similarly, we were unable to account for the instances in which an employer renamed their company, yet stayed at the same address to avoid weightier consequences of violation. This is also a serious problem among these employers and should be studied separately.

Lastly, our analysis implicitly assumes that the census accounts for all migrant workers on H-2A visas and their mobility over time, which it likely does not. It is important to consider that employer work sites may not be located in the same census tract as the housing sites for H-2A workers and

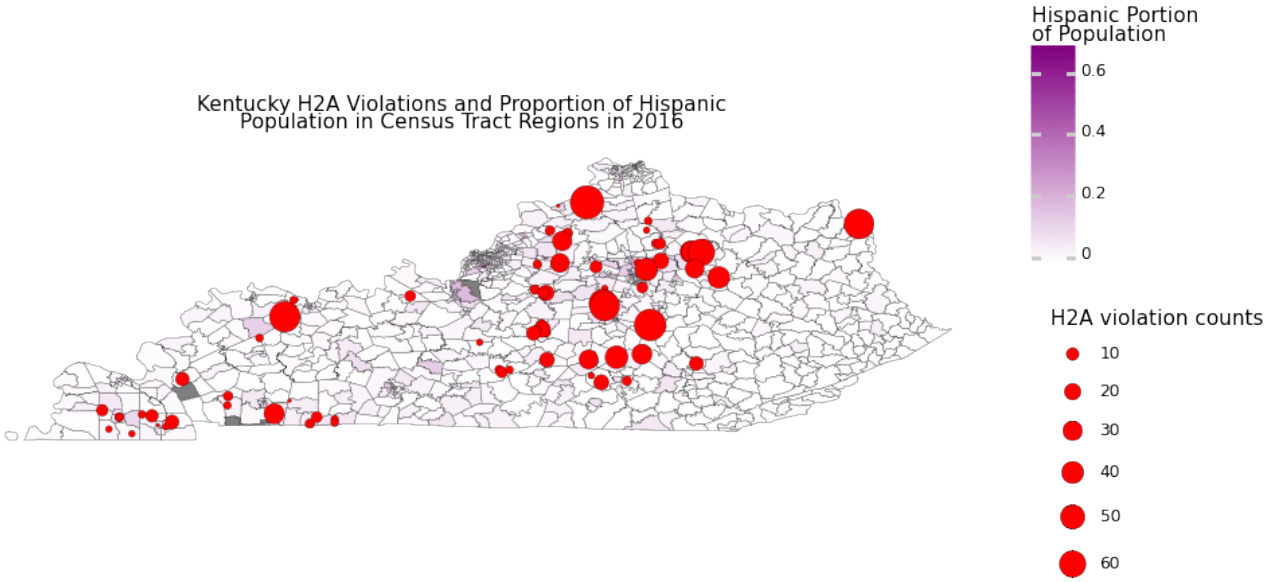

Figure 3: Similar to Figures 1 and 2, this is a graph with a high correlation between the Hispanic population of Kentucky and the H2A violations in a census tract. R^2 = .852

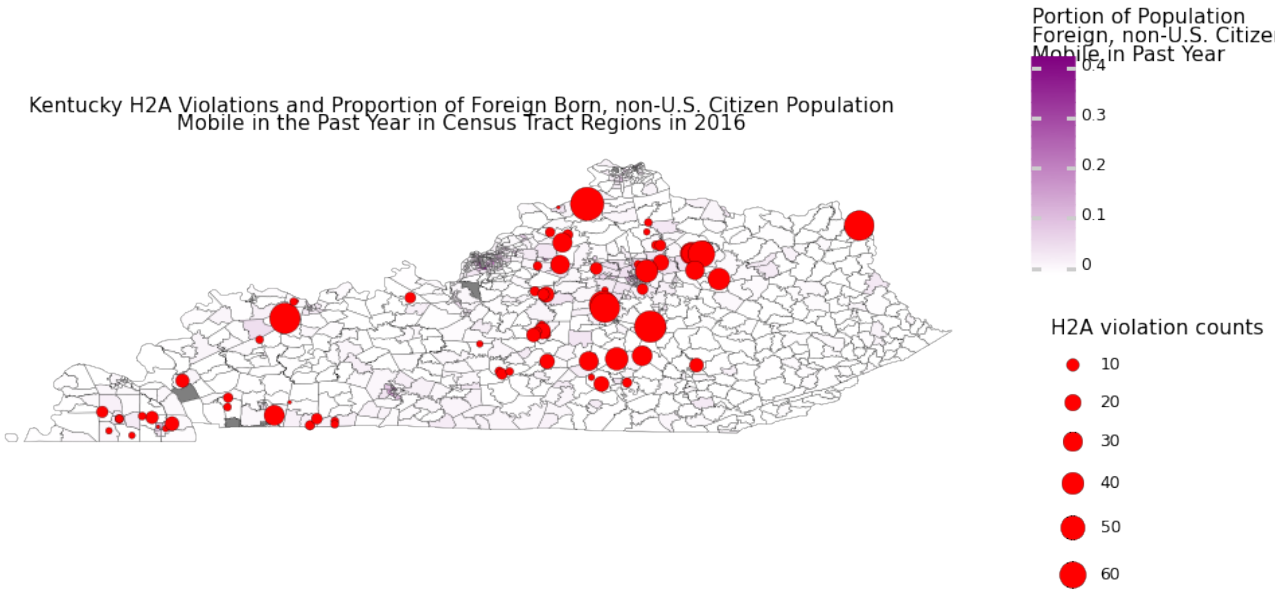

Figure 4: Similar to Figures 1 and 2, this is a graph with a high correlation between the foreign born, non U.S. citizen population of Kentucky and the H2A violations in a census tract. R^2 = .879

that the demographics of the communities surrounding work and housing sites may be important, unobserved variables in the current model.

The findings presented in this paper have important implications for the work of Texas Rio Grande Legal Aid. As shown, there is moderate correlation between the three demographic variables studied and the number of H-2A investigations. Although it was beyond the scope of this paper to investigate more than three demographics from the ACS data, there is significant opportunity for discovering trends using the approach formerly presented. The ACS data includes over 1,000 columns of demographic data on the census tract level, which, if explored further, could provide stronger insight into the types of locations where the H-2A violations are concentrated. This work is important in the ability to predict and prevent employers from violating the H-2A laws, and we hope that this paper can be used in the process of building predictive models to better anticipate the locations of the violations. With the code created, more variables can easily be pulled and more statistical relationships can be investigated. By more thoroughly understanding the characteristics of the tracts where H-2A violations are clustered, TRLA can identify and target these areas for intervention and prevention to protect the H-2A workers.

7 Appendix

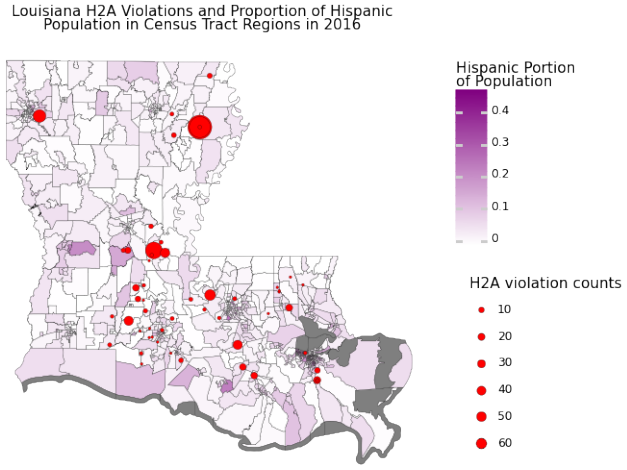

Figure 5: Louisiana Hispanic Population Graph, R^2 = .747

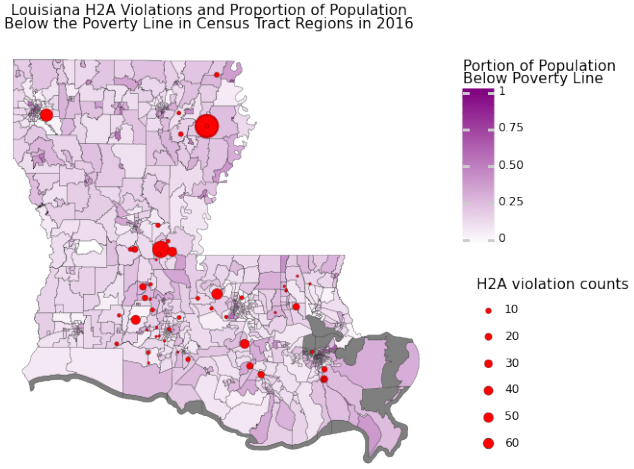

Figure 6: Louisiana Poverty Proportion Graph, R^2 = .440

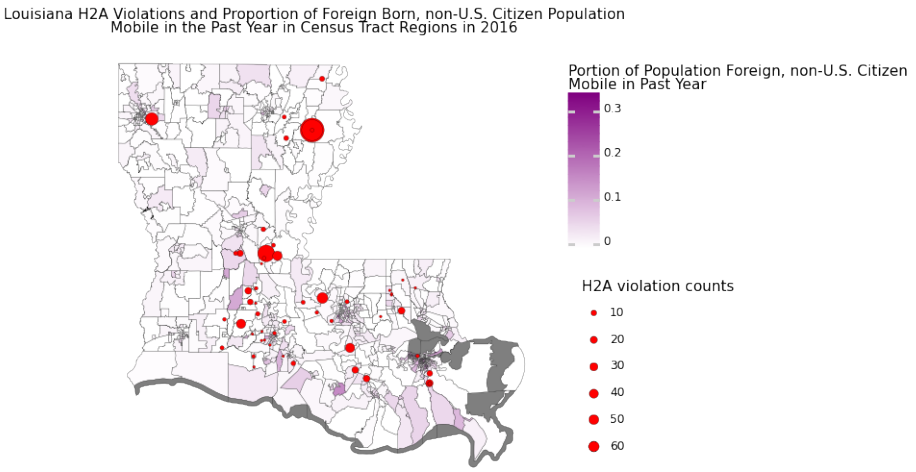

Figure 7: Louisiana Mobility Graph, R^2 = .492

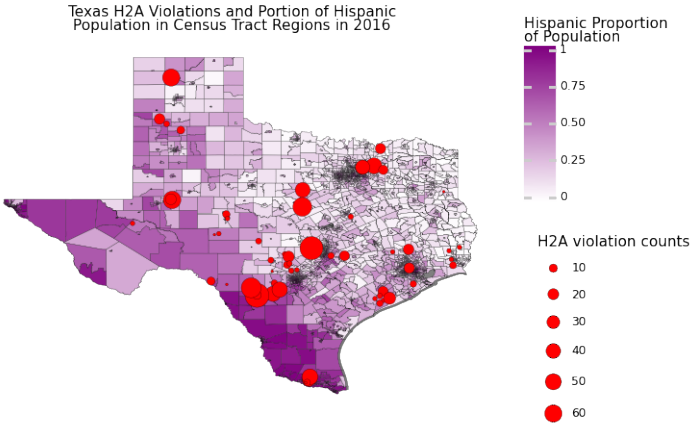

Figure 8: Texas Hispanic Population Graph, R^2 = .799

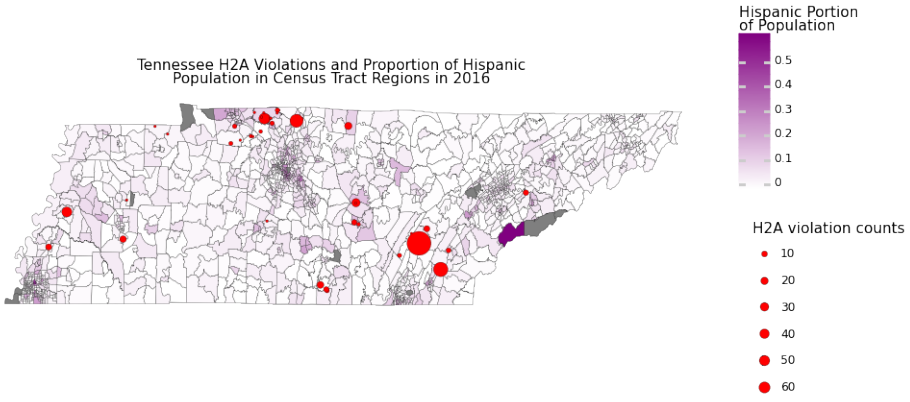

Figure 9: Tennessee Hispanic Population Graph, R^2 = .277

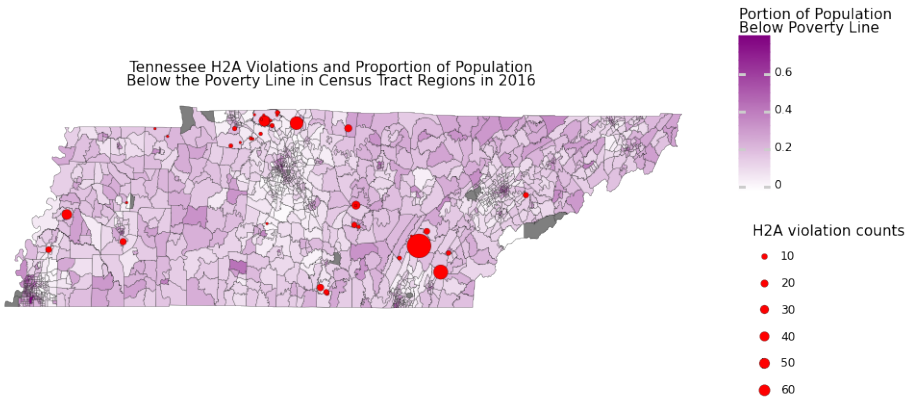

Figure 10: Tennessee Poverty Proportion Graph, R^2 = .563

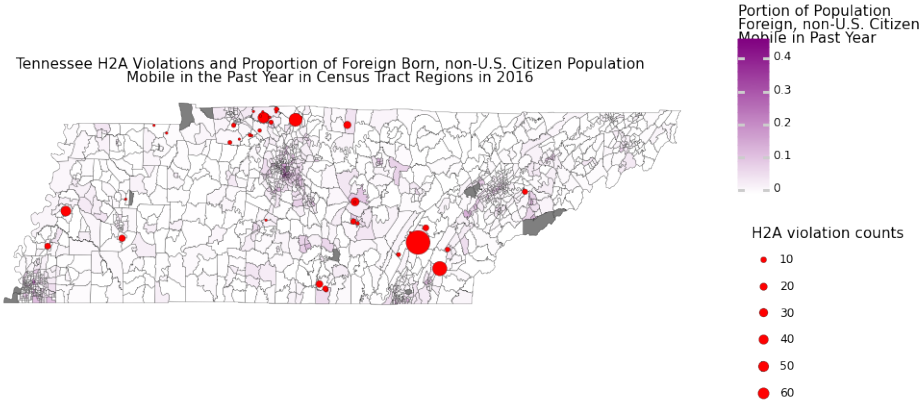

Figure 11: Tennessee Mobility Graph, R^2 = .379

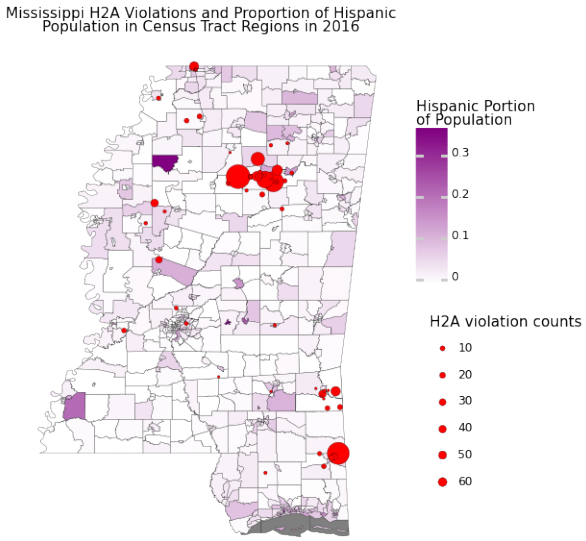

Figure 12: Mississippi Hispanic Population Graph, R^2 = .706

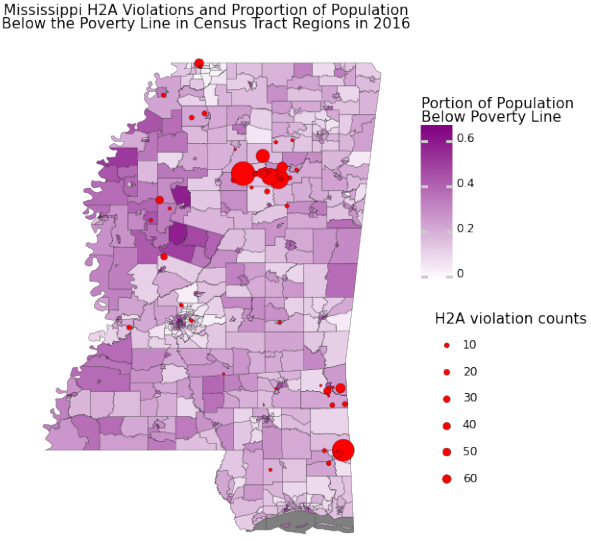

Figure 13: Mississippi Poverty Proportion Graph, R^2 = .517

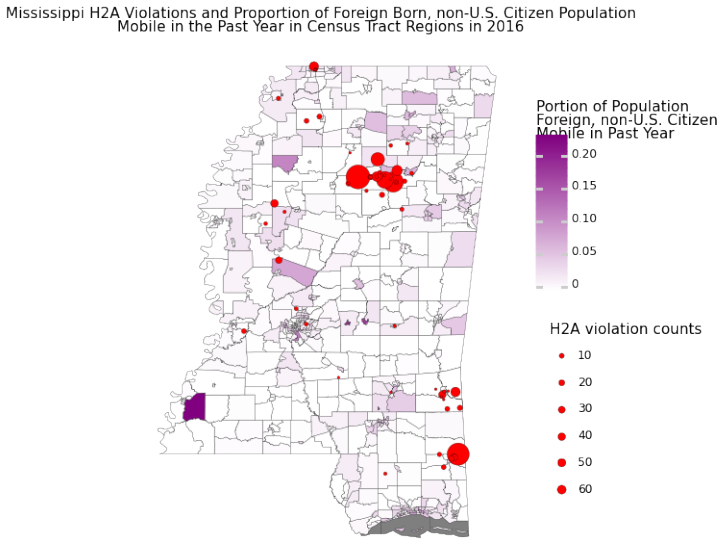

Figure 14: Mississippi Mobility Graph, R^2 = .789

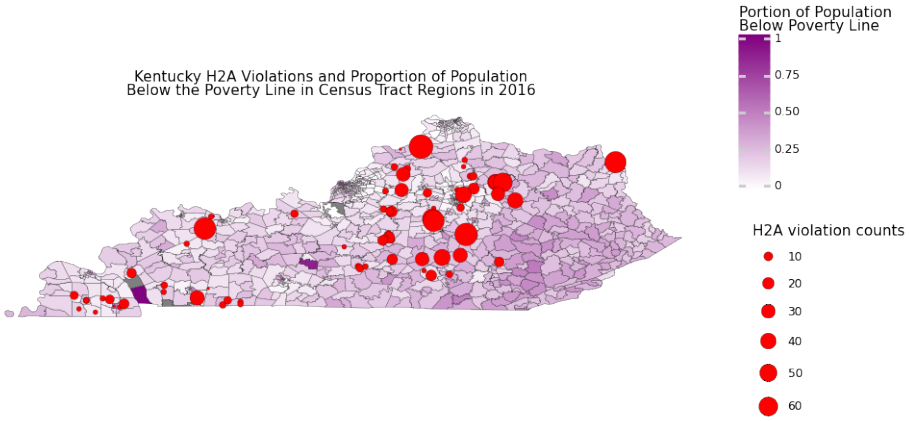

Figure 15: Kentucky Poverty Proportion Graph, R^2 = .738

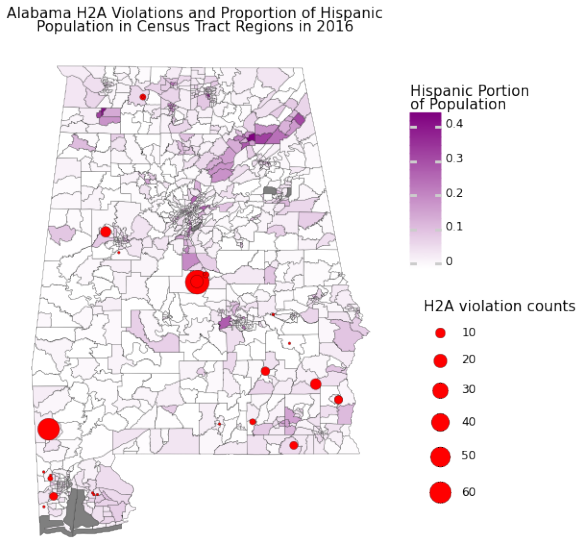

Figure 16: Alabama Hispanic Population Graph, R^2 = .369

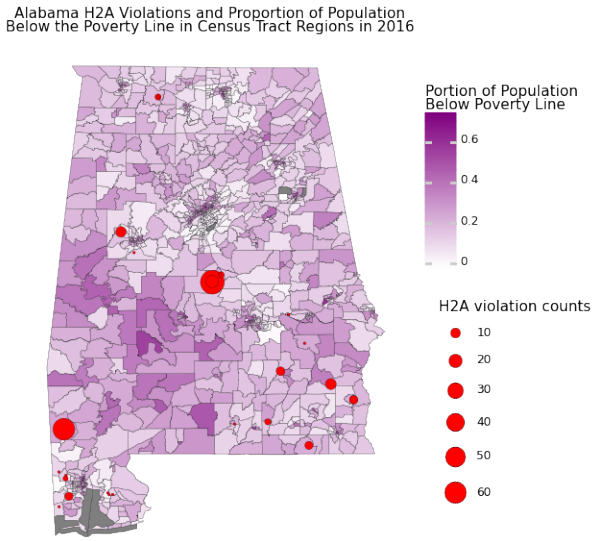

Figure 17: Alabama Poverty Proportion Graph, R^2 = .279

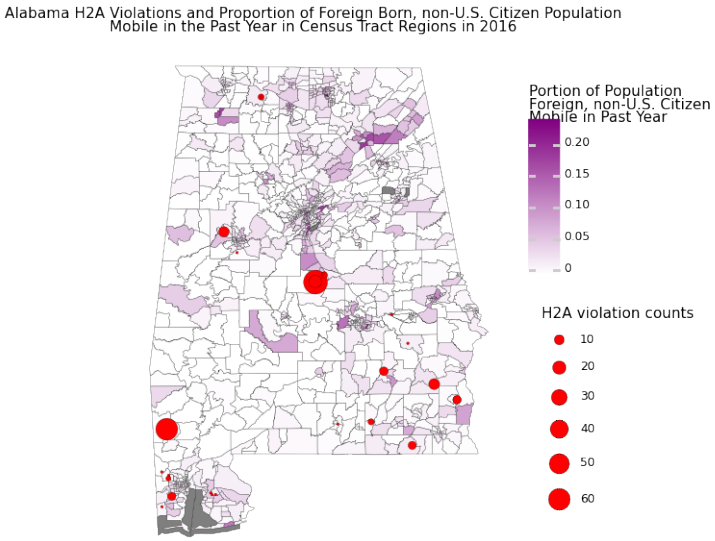

Figure 18: Alabama Mobility Graph, R^2 = .643