The ACS Sandbox

An Interdisciplinary Exploration of the American Community Survey Data

Prepared for ST599 - Big Data Analysis by:
Andrew Bernath
Heather Kitada
Ethan Edwards
Nandhita Narendra

Due date: 4/21/2014

Introduction and Motivation

Statistics is a unique subject. Drawn together by a common interest in data manipulation and interpretation, its practitioners gather from differing backgrounds, each with different sets of interests and expertise. Luckily, many big data projects require this level of diversity to properly explore rich data sets that include a wealth of information from many disciplines. A general survey such as the American Community Survey (ACS) is just this type of data set; a large sandbox in which a number of discoveries may be made. With four members on our analysis team, it is natural to form four different questions to explore:

- Is there a difference in reported commute times between Portland and Los Angeles?
- Is there a spatial relationship in predominant energy type usage across states?
- How do proportions of self-sufficient veterans differ across states?
- How do income levels of immigrants compare to those of US born workers?

This study is intended to summarize key features related to each of these topics.

Results and Findings

A comparison of Los Angeles with Portland highlighted several differences in commuting patterns between these cities. Figure 1 (*below*) shows reported commute times by mode of transportation and percent use by type. The overall mean reported commute times across all modes (represented by the horizontal lines) in Los Angeles are approximately 5 minutes longer than for Portland (29.2 min vs. 24.7 min). Cars, trucks, and vans represent the predominant mode for both cities; however, reported commute times from alternatives to personal vehicles (cycling, walking, subway, and streetcar) in Portland are closer to the overall mean commute times across all modes than the corresponding forms of transportation in LA. Proportions of use for these modes are also higher in Portland, suggesting that as sprawl decreases, people are more likely to seek out alternatives to personal vehicles. An interesting extension of these results would be to compare other cities with known sprawl to those with growth boundaries and compare the trends.

Figure 1: Commute Time Comparison of Los Angeles and Portland by Transportation Mode

	Percent Use by Type	
Transportation Type	Los Angeles	Portland
Bicycle	0.81	2.75
Bus	5.83	5.33
Car/Truck/Van	83.81	79.01
Ferry	0.01	<0.01
Motorcycle	0.29	0.44
Other	0.81	0.58
Railroad	0.25	0.32
Streetcar	0.06	0.70
Subway	0.4	0.64
Taxi	0.05	0.03
Walked	2.82	3.41

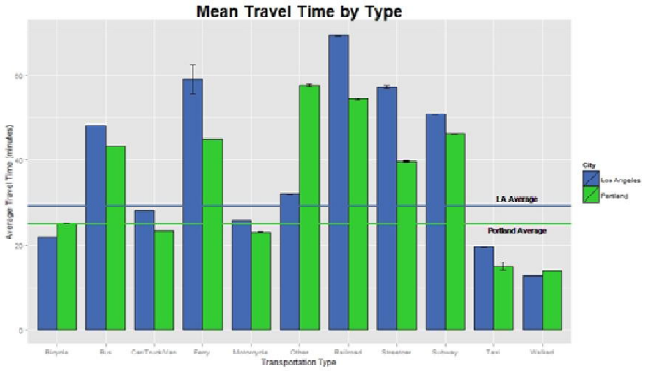
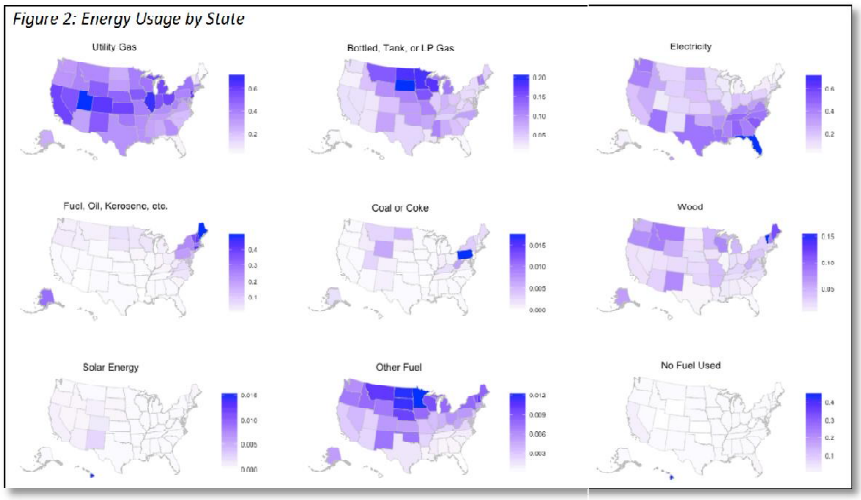


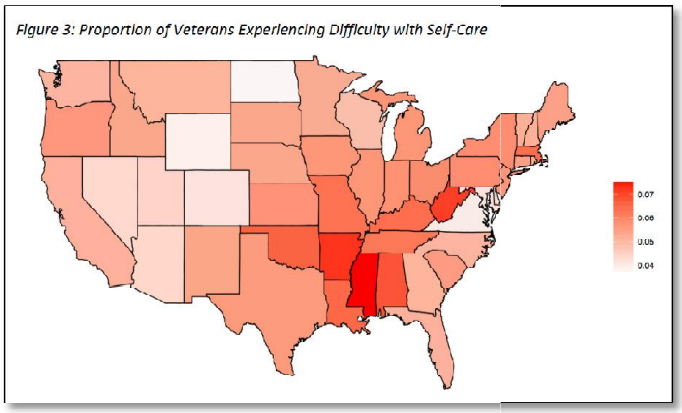
Figure 2 (*right*) illustrates the usage of the energy types represented on the ACS by state. Darker colors represent higher adoption rates. Types of energy usage vary considerably both by state and region and spatial relationships among the states are apparent for nearly every fuel type, each clustering in various areas. Not surprisingly, utility gas shows the highest levels of adoption across all states. Electricity has wide adoption as well, with strongest usage reported in southern states. Residents of central northern states report highest usage of bottled, tank, or liquefied petroleum (LP) gas and "other" fuel sources. Wood is reported most often in the northwestern and

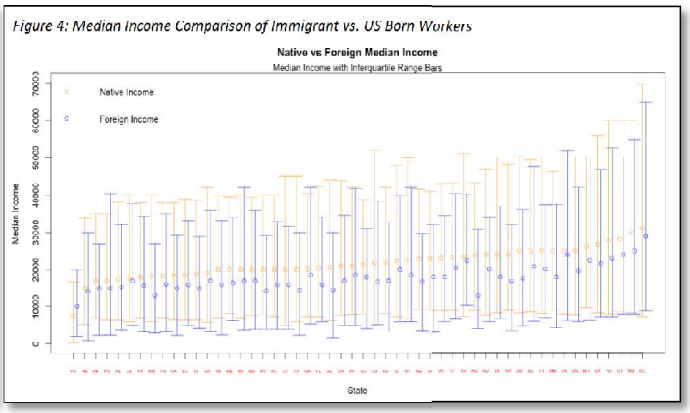
Figure 2: Energy Usage by State



northeastern states. Oil and kerosene are confined nearly entirely to Alaska and New England states. Hawaii leads the US in solar energy adoption (at 1.5% of households) with only light adoption in other regions, primarily the western US.

An examination of the proportion of veterans experiencing difficulty with self-care is presented in Figure 3 (*below, left*). The darker colors in the south represent a higher proportion of veterans experiencing difficulty. Areas with higher proportions tend to cluster together, suggesting that spatial trends are present. Surprisingly, Virginia and Maryland appear to have very low proportions, despite higher incidence in the surrounding states. Investigating differences in availability of veteran care in the lower incidence regions would be a noteworthy opportunity for future studies.





Finally, median incomes were compared between immigrant workers and US born workers across all states. The whisker plot (*Figure 4: above, right*) displays direct comparisons of median incomes for the two groups of workers. Due to the large overlap in interquartile regions for each state, one should be wary in stating that these are significant findings; further testing would be needed to make such a statement. It is still interesting to note; however, that the median income is showing as higher for native born workers than for immigrants in every state (with the exception of Puerto Rico). A good next step would be to actually test the significance of these differences.

Applications and Future Improvements

In order to have more accurate estimates, future studies should incorporate sample weights at various levels (housing, district, state, etc.). Non-response should also be accounted for in a more rigorous examination. There are several useful applications and opportunities that arise from the above summaries, some of which have been previously mentioned. Results from the commuting summary can be used to focus resources on types of transit that are better suited to various amounts of sprawl as contrasted with cities constricted by growth boundaries. Public awareness campaigns regarding alternative forms of energy (such as solar) can be targeted to regions of low adoption. Self care status of veterans can assist with more informed VA healthcare programs. Finally, policy improvements can be made to improve immigrant income levels.

Obstacles

A number of hurdles were encountered while completing the project. In general, several group members experienced difficulties with GitHub not functioning properly. With four people working on four topics, coordination and continuity were a primary focus. A lack of communication surfaced and created a difficult situation for the team. Online tools such as Google Groups and Google Drive helped to alleviate these burdens. In addition to these, several topic specific difficulties arose. For the commute times, the city boundaries were difficult to define as Portland is contained in multiple counties versus LA in a single county. To ensure that a majority of commuters in each city were captured, counties containing the city proper and adjacent cities were included in the study. There was also no clear option for light rail, leaving questions as to which category Portland's light rail was counted in. The energy study arose after a first question regarding internet adoption was discarded due to lack of data prior to 2012. Also, there was some confusion as to the coding of variables in the energy data. Similarly, for the veteran's self-care status, the ACS data dictionary did not define the distinction between "NA" and "Elected not to answer." It was assumed that these were the same. Finally, there was a good deal of difficulty in determining the most appropriate way in which to display the immigration data so as not to be misleading or biased, resulting in a new graph being developed in the last stages of the project.