C-A Rose Capstone Project Proposal

Caryn-Amy Rose

January 1, 2017

I. Problem to solve

What kind of factors predict commute preferences?

This analysis will identify what factors **variables available in the American Community Survey predict an individual's commuting method,** using the following classifications:

1. Car, truck, or van (I will refer to this as "individual vehicle" though it could contain carpoolers)
2. PT-A) Public Transportation - Bus or Trolley Bus (more flexible, infrastructure-wise)
3. PT-B) Public Transportation - Non-Bus (Streetcar, Subway, Ferry, Railroad (less flexible to implement or adjust))
4. Bicycle
5. Walking
6. Motorcycle

The following other commute classification I will not directly study, although they may be involved as predictive factors for other commuters in the household: * Taxi * Not in the labor force (including persons under 16 years; unemployed;.employed, with a job but not at work; Armed.Forces, with a job but not at work) * Not applicable – worked from home * Other commute method

II. Client Decision-making / Impact

Since this analysis will identify predictive factors in commuting choices, community planning boards will be interested in the results. Specific urban/town design planning applications will to some extent depend on specific findings. In general, planning and design such as the following could be be informed by applying these findings to the demographics and development desires of a particular community.

- Bus routes and schedules
- Business locations
- Rail and other transit expansion plans
- Sidewalk and bike lane planning
- Traffic and safety planning (such as lighting installation), for example based on peak times when non-vehicle commuter traffic needs to be accommodated (pedestrians and cyclists)

III. Data to be Used

American Community Survey data.

Available from several sites; I used the files downloadable from https://www.kaggle.com/census/2013-american-community-survey

IV. Brief Outline of Problem-Solving Approach

A. Range of information

The American Community Survey data includes a wide range of data points on persons and households in the US, from marital status, start date of marital status, occupation, education, recent employment history,

and numbers and ages of residents, to household facilities/major appliances, income, and broadband access. People have used it for such different questions as comparing rents regionally or to analyze whether getting a PhD pays off in higher income.

In addition to individual and household profile data, commute specifics such as the following are included. This granularity will help ensure contextualized, meaningful results.

- Means of transportation to work (specifics listed above)
- Travel time to work
- Arrival time to work
- Departure time from work

B. Data Set Considerations

The data itself is composed of two major datasets: person data (> 3.3 million rows) and household data. The person data is associated to household data by serial number.

- 205 person variables (+-, depending how you count/parse some of the information)
- 153 household variables (+-, depending how you count/parse some of the information)

The data dictionary listing variables for both sets is viewable at https://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMS_Data_Dictionary_2009-2013.txt

This data is pretty well organized. It will not require as much cleanup as some sources. The challenge will be in the volume of data and finding the greatest meaning in the very wide range of variables. The data set presents an interesting opportunity in experimenting with a variety of combinations of variable relationships and visualizations.

To pull a sample for running all my preliminary analysis:

1. I randomly selected 184,500 thousand households out of the household data.
2. I then pulled all person data matching those selected households (collecting 391,310 person rows).
3. I will need to incorporate weighting information described at http://www2.census.gov/programs-surveys/acs/tech_docs/pums/ACS2013_PUMS_README.pdf (section VI., Weights in the PUMS, page 5)
4. For speed of initial analysis, I will work with this sample data set of around 10% the total to do my initial hypothesis, analysis, and visualization.
5. After these findings I will apply the most meaningful experiments to the full dataset.

V. Deliverables

Typically, this would include code, along with a paper and/or a slide deck. *To be submitted via Github and in a community presentation.*

- Data
- Sample Analysis Code / RMarkup report (on the sample set; may include supplementary background analysis not included in final/full report)
- Full Analysis / RMarkup report (on the full data set)
- Presentation/slides