December 18, 2018
ISYE 859 UW Madison

Evidence of Commute Patterns by Participant Characteristics in the American Time Use Survey 2017

Introduction

Commuting to and from work is a staple of American life. The time spent commuting is generally recognized to be an activity that is at best neutral and at worst has a negative impact on people's happiness and health. Commute times that are too long can have a negative impact on mental health, which can be exacerbated by the amount of control an individual has over where and when they work (Milner et al 2017). Another study found that an additional 20 minutes of round trip commute time reduced job satisfaction by an amount equivalent to a 19% loss in salary (Chatterjee et al 2017). The same study found that an increase in commute time led to a decrease in leisure time satisfaction.

If commuting is something people would avoid if they could, then longer, more onerous commutes therefore should be undertaken by people who either want to (greater reward) or have to (greater limitations). I am interested in potential biases in commute time. Is a longer commute time a socioeconomic burden or are long commutes associated with more valuable employment? Are the people taking longer commute times more likely to be upper class suburbanites interested in higher pay? Or are people with fewer resources having to take longer commutes because of limited opportunities?

A possible counter to this point of view is the idea that travel time per day is broadly invariant across social and geographic differences. A 1994 study using historical sizes of settlements suggests that over time human settlements have formed in patterns based on an average 30 minute travel time. As faster modes of transport have been introduced, average travel per day has increased in distance but not in time (Marchetti 1994). This ideas suggests that on average there is a large amount of similarity in commutes, and that we should not expect to see strong socioeconomic patterns.

To distinguish between these divergent views, I look at relationships between commute time and other respondent characteristics, such as income, age, and time spent with family. By applying clustering and predictive algorithms, I may be able to learn about patterns in commute times. For example, is a longer commute time a socioeconomic burden or are long commutes associated with more valuable employment?

Looking at a daily measure of how Americans split their time, I did not find a strong relationship between commute time and age, income, time spent with family, friends, or children. This study supports strong similarities in commutes times across groups, but the nature of the data used has limitations which suggest the need for further study.

Data

This study uses the 2017 American Time Use Survey Data, specifically the activity, roster, and respondent files. Please see the associated file “commutes_final_integration.html” for R analysis details and some additional explanatory figures.

The American Time Use Survey (ATUS) has been run since 2003 by the US Census Bureau. The Bureau fields the respondents from individuals who have been reliable participants in Current Population Survey (CPS), which is a monthly survey used to estimate unemployment. The 2017 data has slightly over 10,000 respondents with about 2000 participants who have provided commute data. The survey includes self-reported information on work, leisure, travel, child and eldercare, and other activities.

Commute characteristics
To look specifically at commute times, I filtered the data to activities classified as “travel for work” which identified the type of vehicle used (bus, train, walking, car). This brought the data set down 5400 instances of one way commutes in the data, the majority by car, representing 2800 participants (figure 1).

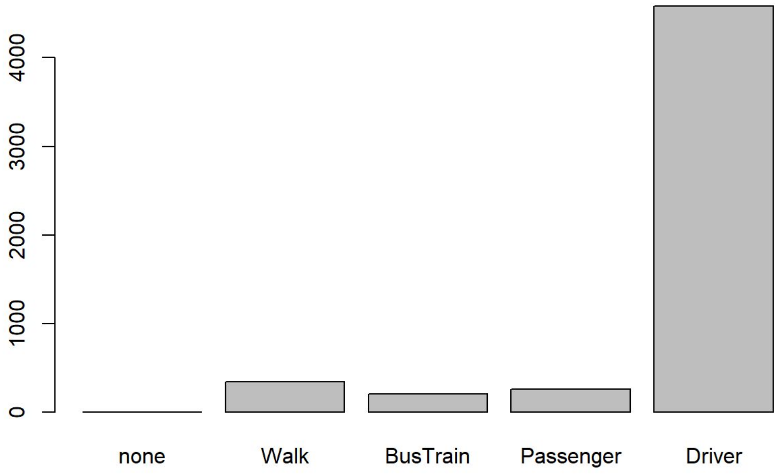

Figure 1. Counts of commute types.

Because the data was not gathered for the purpose of examining commute patterns, attributes describing the commute were limited. Commute attributes in the dataset were limited to start time, end time, duration, and location (e.g. bus, car…).

Participant characteristics

To tie individual commute trips to participants, I joined the activity data to the respondent data. Again, the dataset was not gathered with commutes in mind so there were some limitations to the characteristics available. Attributes included in the dataset were age, hours worked per week, spouse hours worked per week, number of children, hourly wage, time alone, time with children, time with family, time with friends. Additional attributes had to be discarded because of poor data coverage. This filtering step brought the dataset down to about 1300 participants.

Analysis: Clustering and Predictive models

First, I clustered commutes by characteristics (time, duration, type). Next, I compiled commute time per day by participant and used a gradient boosted decision tree algorithm to predict total commute time from respondent characteristics. No other transformations were made to the data.

Clustering

In this analysis step, I used unsupervised learning uncover patterns in people's commute habits. I used a kmeans clustering method because it is a good all-purpose algorithm, and the data does not appear to be so unusual as to need a different clustering algorithm. When testing additional clustering methods, there was not a strong difference in performance. Metrics about quality of clustering suggested that 2 clusters was the ideal number for this dataset, but because that limited the ability to detect variance in the data, I increased the number of clusters to 6 which was able to highlight some differences in commute type. In general, the cluster algorithm found passenger commutes to be the most unlike other commutes. As cluster number increased, the algorithm tended to break down car and walking commutes by time of day. Because commute durations did seem to be quite similar, there was not a discernible difference in start to stop time as a clustering criteria. Also, a morning and evening commute can be detected in the histogram above the graph (figure 2).

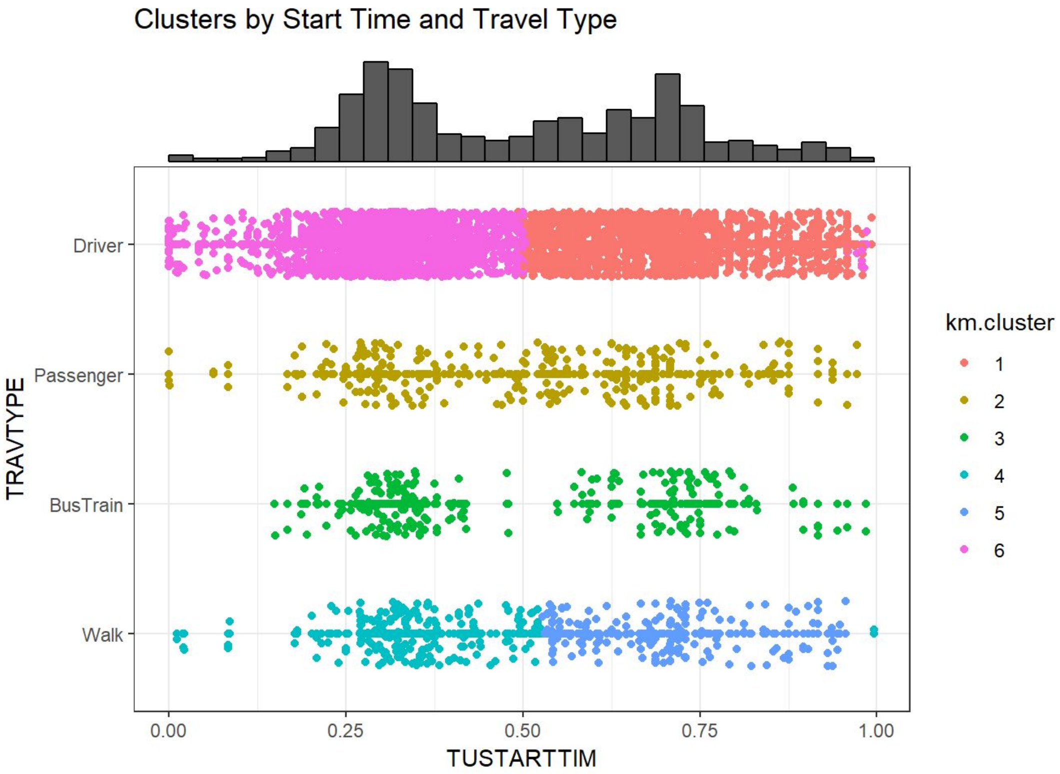

Figure 2. Kmeans clusters by travel type and start time (percentage of 24 hr period starting at 4 a.m).

If we graph these clusters as functions of participant age and hourly wage, we can see that no pattern emerges (figure 3).

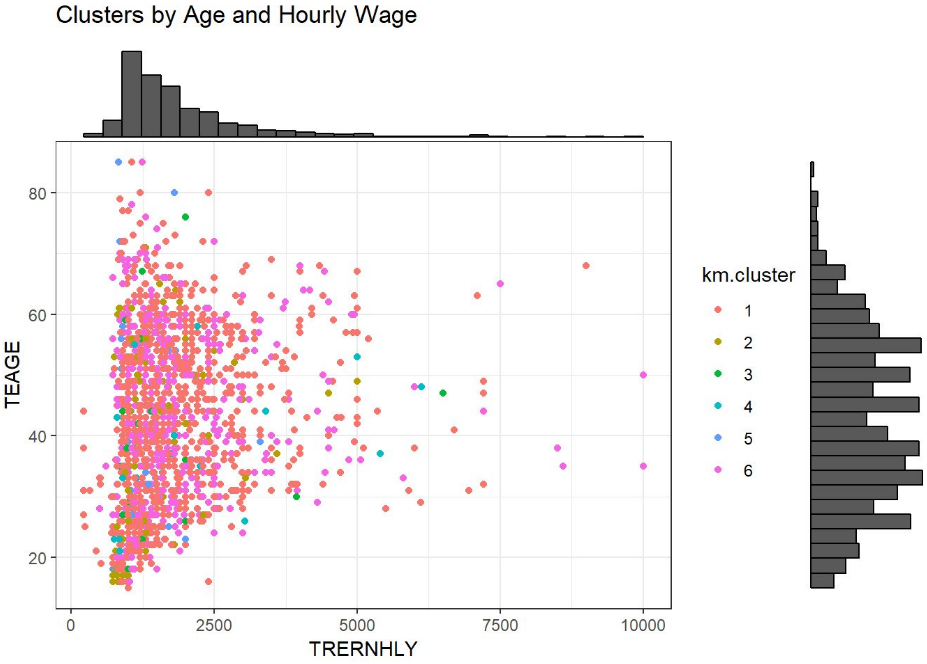

Figure 3. K-means clusters plotted as a function of age and hourly wage.

The next step is to put all the participant characteristics into a supervised learning algorithm and see if we have any predictive power.

Prediction

In my second analysis step, I used supervised learning to predict total daily commute duration for each respondent. As mentioned in the previous section, attributes describing participants were limited by data coverage and survey design, which was not intended for this purpose. That said, attributes related to quality of life (time spent alone, with family, friends, and children, hours worked per week) and socioeconomic status (hourly wage) were included in the dataset.

To predict total commute times, I used the caret package in R to run a linear model and a gradient boosted decision tree to see how well available attributes could predict total commute times. Both methods made very weak predictions. The linear model prediction had an R^2 of 3.2%. The gradient boosting method (xgboost) had an R^2 of 4.9%.

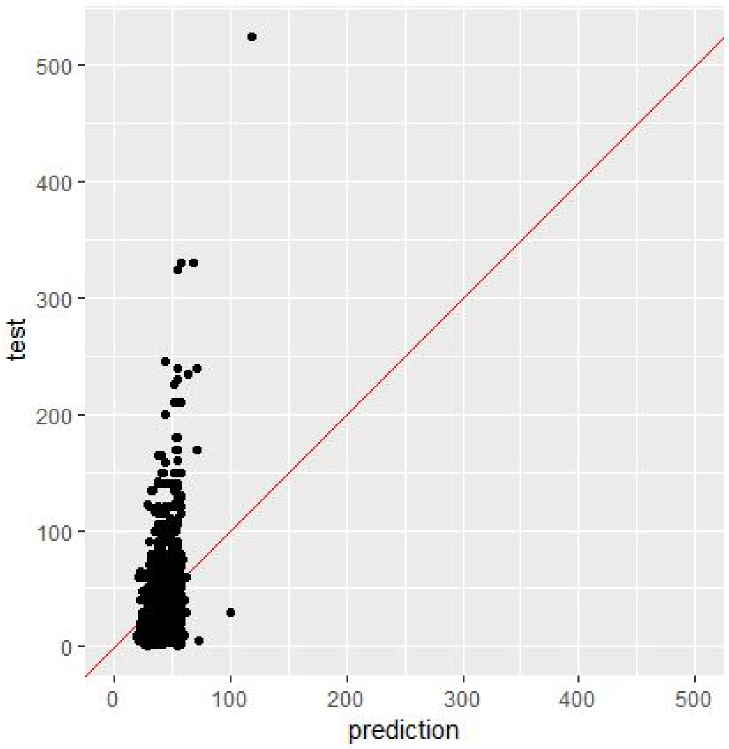

Figure 4. Predictions of the gradient boosting algorithm. A perfect prediction would fall on the red line. Variance in the independent variables did not predict variable in the dependent variable. Predictions tended to have less variance and magnitude than the measured data.

Variable importance also varied significantly between models, despite both being similarly weak in statistical power, another indication of the lack of explanatory power in the models (figure 5). The top 3 most important variables are completely different between models. Also, ‘time alone’ may actually include commute time for participants who drove themselves, hence its explanatory power may lie in redundancy with the dependent variable, total commute time.

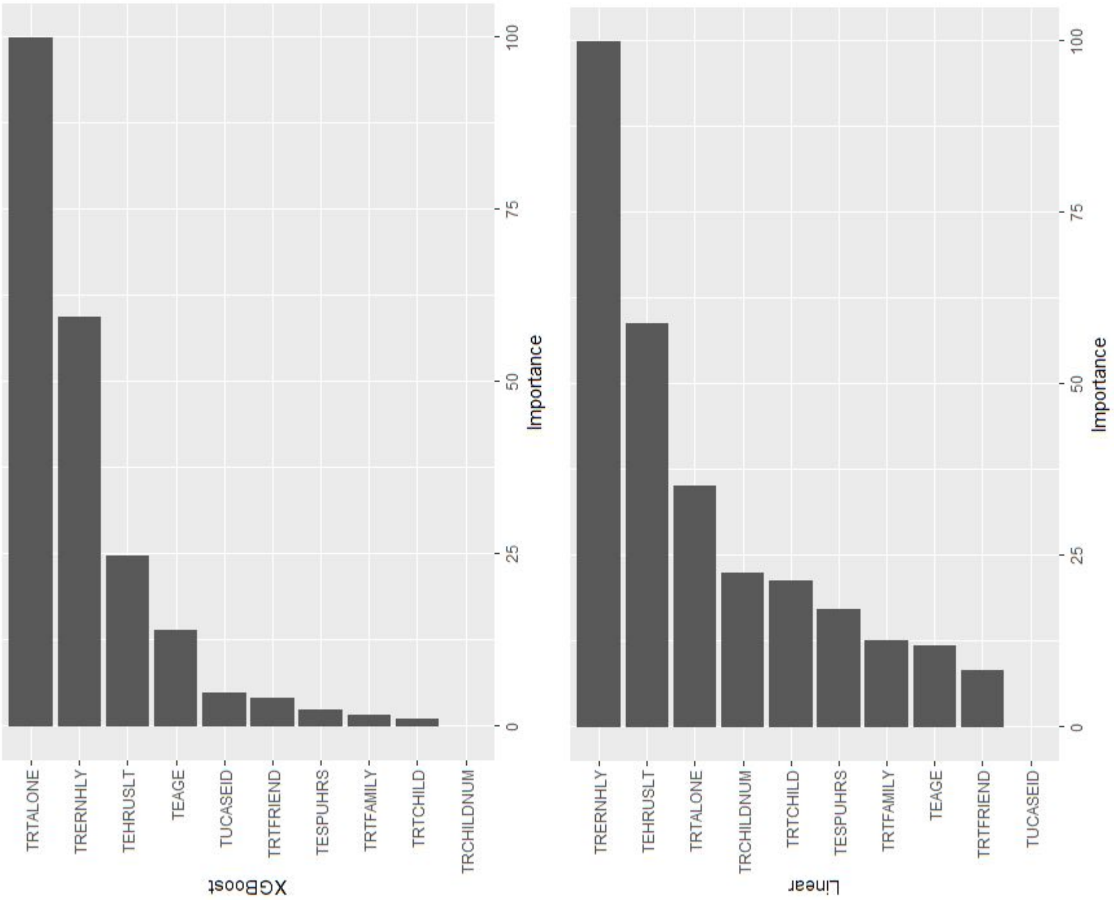

Figure 5. Variable importance charts for (a) XGBoost and (b) Linear model

Age, hourly wage, time spent with others are not strong predictors of commute time. It should be noted that for time based variables like time spent with children, the dataset can only see activities that coexist in one day. Therefore, the predictive model asks, effectively: for Americans, are days with longer commute times likely differ in time spent with friends, family, or children from days with shorter commute times? This model offers a prediction of an average day. That prediction is: in this data set, days with longer commute times will not differ strongly from days with shorter commute times regarding time spent with children, friends, or family or time spent alone. There was also not strong predictive power in multi-day spanning attributes like age, hours worked per week, number of children or hourly wage.

Further details of the R analysis can be found in the associate R markdown output file.

Weaknesses

A challenge with using this algorithm is that not many characteristics of commuting were gathered in this data set. There is very limited information about the commute outside of when it started and stopped. The vast majority of the data is bucketed into "car/motorcyle (driver)" which means it will be difficult to detect variation in behavior. As such, the conclusions that can be drawn from the data may be limited.

This model has some weakness in validity and generalizability. The American Time Use survey was not designed to measure commute details, so an assumption has been made that "work related travel" indicates a daily commute for the majority of participants. Also, problems with the reporting of location during the commute mean that of the 10,000 or so participants, only a few thousand were included in a nonrandom manner. However the sample size is still relatively large and comes from a representative sample of the US population, so that patterns found here would be a good starting point for further investigation.

Discussion and Next Steps

The most interesting part of the model is that no strong pattern emerges. In this dataset, the independent variables included in the model (age, hourly wage, with family, time alone, number of children) do not strongly predict how long an individual commutes per day. This may suggest that people in different age and income groups generally agree on an acceptable commute time, and either choose their home based on their work or their work based on their home. In this way, the data are in line with Marchetti's 1994 study which found broad agreement within the human race over time about preferred travel time per day.

However, this study does rule out the possible of bias in commute times. The impact of such biases may be felt in the margins, not among the average participants, which this study was not well equipped to examine, due to limited data points and the non-random nature of the filtering.

Text Analysis with Machine Learning

Furthermore, many variables in this dataset remain unexplored, however, and there may be undiscovered relationships. One potential next step, would be to apply text analysis that identifies topics across documents, known as structural topic modeling. The activity record of a participant in the dataset is similar to a piece of text in that there is a large number of possible activities (words) but each participant (document) contains only a small subset.There are 397 possible activity categories in the data. 311 activities average less than 1 minute per participant, likely because very few participants reported doing the activity.

Applying structural topic modeling to the activity dataset would allow a machine learning algorithm to assign common topics to participants across all possible activities, not just the ‘work related travel’ used in this study. By doing patterns of associated activities may emerge in ways that uncover interesting relationships.

In the meantime, this study offers an example of the ways in which commute patterns cross socioeconomic and lifestyle boundaries.

Works Cited

Milner, Allison, Hannah Badland, Anne Kavanagh, Anthony D. LaMontagne (2017). Time Spent Commuting to Work and Mental Health: Evidence From 13 Waves of an Australian Cohort Study, American Journal of Epidemiology. Volume 186, Issue 6, 15, Pages 659–667. https://doi.org/10.1093/aje/kww243

Chatterjee, K., Clark, B., Martin, A. & Davis, A. (2017). The Commuting and Wellbeing Study: Understanding the Impact of Commuting on People’s Lives. UWE Bristol, UK. https://travelbehaviour.files.wordpress.com/2017/10/caw-summaryreport-onlineedition.pdf

Marchetti, C. Anthropological Invariants in Travel Behavior (1994). TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE. Volume 47, Pages 75-88. https://doi.org/10.1016/0040-1625(94)90041-8