**Predicting Public Transportation Demand Using Household and Neighborhood Characteristics**

Alexa Canaan

Colgate University

**Predicting Public Transportation Demand Using Household and Neighborhood Characteristics**

Abstract – *This paper aims to predict public transportation demand by comparing the household and neighborhood characteristics of those who do and do not use public transportation. Data from the 2013 American Housing Survey is analyzed to compare public transportation usage across a representative sample of households in the United States. Both OLS and logit regression models found that distance from public transportation is a statistically significant factor in determining public transportation usage. More coming soon…*

**Key Words:** Public Transportation, OLS, Logit Regression, LASSO Regression, Public Economics, *more coming soon*

**JEL Codes:** *coming soon…*

1. **Introduction and Motivation**

Public transportation has the ability to revolutionize a geographic area by increasing travel speeds, decreasing commute times, facilitating labor market matching, and providing more opportunities for leisure and consumption (Chin, Kahn, Moon, 2017, p.3). For individuals, the ability to reach necessary services and spatial opportunities is a key quality of life metric that can be improved with a range of transportation options, public and private (Karou et al., 2014, p.1). Not only that, but with traffic congestion being a source of economic cost for most industries and sectors, increasing the general use of public transportation can ease traffic flows and significantly increase productivity for individuals and benefit business (Weisbrod, Vary, Treyz, 2003, p.98). However, this is only true to the extent that public transportation is available to a population that will use it. For people who live far away from public transportation to the point that it is inaccessible, they cannot take advantage of its benefits even though they may choose to do so if it was accessible. My motivation for conducting this analysis is to predict public transportation usage across America using household and neighborhood characteristics. Variables of interest including distance to public transportation and metro status for households can be used to gauge accessibility to public transportation which is a key component of whether or not individuals actually opt to use it.

This paper will supplement existing research by examining the effect of accessibility on demand within the context of public transportation. Most research thus far has been concentrated on increasing use of public transportation by changing people’s attitudes towards public transportation. However, this research relies on the assumption that we can only increase ridership among populations where public transportation is already available. We should also consider accessibility of public transportation as a key component of its usage. If accessibility is a significant determinant of public transportation use, then policy makers may want to consider investing in making public transportation more accessible rather than trying to change their constituents’ attitudes. This paper specifically will aim to develop a machine learning model that can be improved upon and used in the future to predict demand of public transportation based on household and neighborhood characteristics of household observations. Ideally, this model will allow researchers to identify inefficiencies in access to public transportation. These inefficiencies occur when demand does not match accessibility, such as when a neighborhood has high demand for public transportation, but there are no stops nearby or when a neighborhood has little to no demand for public transportation, yet there are potentially multiple stops nearby. The predictive capability of the model will allow for policy makers and urban planners to better allocate infrastructures to communities across the United States based on whether or not their accessibility to these public goods correlates to them using said goods.

1. **Background Literature and Conceptual Framework**

Public transportation demand brings together several different avenues of economic thought, specifically discrete choice theory, estimating demand, and the transportation sector. Predicting demand has become increasingly important, especially within the public transportation sector, because it allows policy makers to optimize the transportation network holistically (Palacio, 2018, p.3). Optimizing transportation networks can have important social consequences because a lack of access to a range of transportation options is a source of inequality. Over time, “the spatial growth of urban areas and the decentralization of employment and facilities have made it harder for people without access to a car to make the daily commute and to take advantage of distributed retail and leisure opportunities” (Karou and Hull, 2014, p.1). This pattern systematically limits opportunities for people who do not live near public transportation and cannot afford a car or can only afford to live within their means by limiting their gas consumption. Therefore, to support sustainable urban growth for all people, it has become increasingly important for urban planners to make informed decisions on land use and transport decision-making (Karou and Hull, 2014, p.2). Especially with public transportation infrastructures taking large amounts of time and resources to develop, making sure that these services will be utilized by populations who rely on them for employment, educational opportunities, and affordable consumption is key to supporting sustainable and equitable urban growth. This paper will hopefully emphasize the importance of accessibility in decision-making with regards to public transportation use but can also serve as a model for predicting demand more generally.

Another social issue that access to public transportation can ameliorate is global climate change. Reducing greenhouse gas emissions is key to lessening the effects of climate change, yet heavy reliance on driving cars makes this more difficult (Heath and Gifford, 2002, p.2154). Therefore, much public transportation usage research has been focused on how to get people to switch over from using their cars to public transportation, specifically focusing on people’s attitudes towards public transportation and their concern with the environment. Borhan et al. (2014) studied travelers’ willingness to use public transportation in Putrajaya, Malaysia by focusing on a wide range of variables from service quality, environmental impact, attitude, and behavior intention, with the specific intention of trying to find ways to better public transportation to increase its usage. They found that environmental impact was important for choosing to use public transportation, but so were service quality and general attitude were also important determinants (Borhan et al., 2014, p.5). In fact, environmental impact alone had no significant direct impact on behavior intention to use public transport (Borhan et al., 2014, p.6). It was found though, that providing adequate information about public transport routes are directly related with public transportation usage, meaning that accessibility is an important determinant of its use (Borhan et al., 2014, p.6).

Heath and Gifford (2002) performed a similar analysis employing the theory of planned behavior (TPB) to examine psychological factors motivating public transportation in western Canada. The TPB theorizes behavior is derived from intention, which is an antecedent of attitudes, norms, perceived behavioral control (Heath and Gifford, 2002, pp. 2155-2156). The authors added to the original TPB model by incorporating variables indicating intentions, social norms, and environmental concerns and values to predict public transportation usage (Heath and Gifford, 2002, pp.2157-2159). Specifically, in regard to environmental concerns, they found that pro-environmental behavior is subject to many contextual constraints out of the control of individuals (Heath and Gifford, 2002, p. 2176). This means that while people would opt to use public transportation and simultaneously save the environment, sometimes this is not possible because there is no public transportation that is accessible to them. Again though, it was found that environmental concerns alone cannot convince an individual to use public transportation over their own vehicle (Heath and Gifford, 2002, p.2176).

Disregarding individual attitudes about environmental concern and focusing on personal preferences about public transportation more generally, other research finds that again, many variables about personal preference need to interact to make public transportation a more attractive option than a personal car. Van Oort et al. (2015) focuses on determinants of ridership in the Hague such as comfort and capacity. Using smart card data, they created a generalized cost function measuring the effects of crowding on individuals such as denied boarding and vehicle preferences (van Oort et al., 2015, p.5). They found that incorporating comfort and capacity along with other demand determinants improved the predictability of a public transportation demand modelling framework (van Oort et al., 2015, p.12). Overall, it has been found that many variables need to interact in order to marginally increase use of public transportation, even when it may already be accessible to a population. However, making public transportation accessible to begin with is important in establishing a base level of demand.

By focusing on accessibility, this paper contributes to information on public transportation demand by identifying a base level or long-term level of demand for public transportation that could support research that aims to identify short-term levels of demand. Palacio’s (2018) paper using smart card data, for example, aimed to predict daily and hourly demand of public transportation by looking at variables indicating peak travel hours, whether or not it was a workday, the season, and daily weather. Menon and Lee’s (2017) analysis using bus arrival data from an Australian metropolitan area also planned to research short-term public transportation demand, within 5 to 10 minutes of a bus’s arrival at a stop. Short-term levels of demand can also be nonhabitual, like in the case of special events. For example, Pereira et al. (2015) use web mining techniques to identify nonhabitual, disruptive levels of public transportation demand due to planned special events in Singapore. Large-scale events can be especially disruptive to a public transportation system because these systems are built based on habitual behavior and therefore usually rely on special coordination efforts on days that these large-scale events occur (Pereira et al., 2015, p.2). The researchers found that event information found on the web using APIs improves the quality of public transportation demand predictions under special event scenarios (Pereira et al., 2015, p.24). This research allows the government to refine public transportation infrastructure that is already in place by increasing or decreasing the frequency of buses and trains on route at different time intervals, as well as tweaking scheduling services if necessary. However, this paper will aim to predict whether or not certain areas require a large investment in infrastructure for public transportation using accessibility of public transportation as a determining factor of usage.

This paper adds to current literature by utilizing data from all population densities in the United States from urban centers to rural areas. Papers such as Omrani (2015), Palacio (2018), Chin et al. (2015), Pereira et al. (2015), and van Oort et al. (2015) all focus solely on urban centers. While data on public transportation is usually only available for metro areas because this is where these systems are placed, researching these systems in a macro context may expose a need for public transportation in areas that are not considered typical areas for this kind of infrastructure.

Recent research has focused on using machine learning to predict discrete choice and demand specifically for public transportation. Since most machine learning techniques are relatively new, many papers related to prediction have compared the accuracy of different techniques to justify their results (Omrani, 2015, p.841). This is also beneficial to economics and research as whole as it contributes to the goal of identifying what technique is best for making predictions (Kleinberg et al., 2015).

1. **Data Description and Analysis**

I use data from the American Housing Survey, a longitudinal housing unit survey, to develop my predictive model with all computations and graphics performed in the R programming language(www.r-project.org). The American Housing Survey is a comprehensive survey sponsored by the Department of Housing and Urban Development and conducted by the U.S. Census Bureau biennially in odd-number years, spanning from 1973 to 2017 between May and September. The same housing units from all 50 states and the District of Columbia are surveyed every other year until new samples are drawn, allowing for analysis of households over time. The goal of the data set is to provide timely information on the quality and cost of housing in the United States and American metropolitan areas using the participating housing units, chosen to represent all housing units in the United States. It is used by policymakers to make decisions about housing for all demographics in America (US Census Bureau, 2020).

Each observation in the data set is defined as a “housing unit” or any house, townhouse, apartment, mobile home or trailer, single room, group of rooms, or other location that is occupied as separate living quarters, or if vacant, is intended for occupancy as separate living quarters (US Census Bureau, 2020). The survey is conducted using computer-assisted personal interviewing using laptops. Data is collected from two types of respondents: occupied housing units and unoccupied housing units. Data collected from an occupied housing unit is defined by the US Census Bureau as “A household respondent, who must be a knowledgeable household member 16 years of age or over, provides information on the unit, the household composition, and income” (US Census Bureau, 2020). Data collected from a vacant housing unit is defined by the US Census Bureau as coming from a respondent such as “a landlord, owner, real estate agent, or knowledgeable neighbor” who can provide data on the unit (US Census Bureau, 2020).

This particular data set is applicable to the research question because it has been used by policy makers to plan community development such as infrastructure. I will specifically focus on data from 2013 due to its large range of questions covering public transportation. There were approximately 84,400 sampled housing units with a supplemental sample of 15,533 housing units in the Chicago, Detroit, New York City, Northern New Jersey, and Philadelphia metropolitan areas (US Census Bureau, 2013). Out of the 84,400 sampled housing units, 2,715 were ineligible because the unit no longer existed or because it did not meet the definition of a housing unit (US Census Bureau, 2013). Further, 10,000 units had no response after repeat visits or refused to be interviewed which led to an overall response rate of 86% (US Census Bureau, 2013).

The data does come with its own set of limitations in terms of incomplete data, wrong answers, and sampling variability, much of which is unavoidable. For example, incomplete and missing data are adjusted by assuming that the respondents are similar to its peer households that did report data, assigning them an average value instead of ignoring the missing data (US Census Bureau, 2013). The dataset is not adjusted for wrong answers and does not estimate the size of these errors (US Census Bureau, 2013). Sampling errors are accounted for by the US Census Bureau by proving standard error tables and allowing users to calculate their own sampling errors, although these calculations may vary by user (US Census Bureau, 2013). Recognizing the limitations of the data is important because while they are unavoidable, our ability to take the findings out of context is not. Recognizing that these observations can be flawed is critical to our treatment of the findings as estimations.

In order to learn more about public transportation usage in America, several categorical variables were utilized to create my simplified model. The main variable of interest is PTPUBTRN, recoded to be an indicator variable that is 1 if the respondent uses public transportation and 0 if the respondent does not. METRO3 is a variable that indicates whether or not a person lives in a city center (1), metropolitan area (2), or rural area (3). CARS and TRUCKS indicate the number of cars and trucks per household, respectively. ZINC2 is a numeric variable that represents total household income. The variables PTDISPUB, PTDISBUS, PTDISRAIL, PTDISSHUT, and PTDISSUB represent how far each household is from a general public transportation stop, bus stop, rail stop, shuttle stop, and subway stop, respectively. The distances from public transportation stops for each household are measured on a gradient, with 1 representing less than ¼ mile, 2 representing ¼ to less than ½ mile, 3 representing ½ to less than 1 mile, and 4 representing 1 mile or more. Additional variables that are included in the model accounting for neighborhood effects are WNTRAN, WNAMEN, and WNJOB, which indicate with 1 if the respondent moved to their neighborhood to be closer to public transportation, amenities, or their job, respectively and 0 if they did not.

Before performing any regression analysis, it is important to examine the data and make sure it is readable in R. For the sake of being concise, it is necessary to remove factor variables that have only one level, because this means that there is no difference in condition and all observations share the same value for said variable. When doing this in R, the number of variables decreased from 4114 to 3717.

1. **Regression Models and Estimation Methods**
2. **Regression Results and Analysis**

While I do not have any finalized results for my paper, there are some systematic errors that need to be addressed regardless of how I choose to clean and analyze the data. Those ideas are addressed below. I expect to find other errors as I continue with my analysis.

It is important to recognize that due to the nature of the dataset, some groups were better represented than others. The American Housing Survey is supposed to provide a representative sample of the United States based on housing characteristics, but this means that certain populations are given less weight than others because they make up less of the United States population.

There is possible measurement error in regard to measuring distance from public transportation stops to households in our OLS and logit models. Since the distance to public transportation stops are grouped in quarter mile increments and then in a group of greater than 1 mile away from public transportation, we are potentially both underestimating and overestimating how far away households are from public transportation stops. These distance gradients could also mean that there is attenuation bias in our coefficients, meaning the estimate coefficient is closer to zero than it should be.

1. **Summary and Conclusions**

Other topics that beg further questioning from this paper include other household characteristics and public transportation use, such as whether those who rent are more likely to use public transportation than those who own property. Not only that, but investigating topics related to accessibility such as neighborhood safety could help identify what factors bolster and hinder accessibility. Applying the model to the accessibility of other resources more generally like food security could develop a more holistic model of the relationship between accessibility and demand that could be used by policy makers in their general decision making.

1. **References**

**(APA Style)**

Borhan, M. N., Syamsunur, D., Mohd Akhir, N., Yazid, M., Razuhanafi, M., Ismail, A., & Rahmat, R. A. (2014). Predicting the use of public transportation: a case study from Putrajaya, Malaysia. *The Scientific World Journal*, *2014*.

Chin, S., Kahn, M. E., & Moon, H. R. (2017). Estimating the Gains from New Rail Transit Investment: A Machine Learning Tree Approach. *NBER Working Paper Series*, 1–49. doi: 10.3386/w23326

Heath, Y., & Gifford, R. (2002). Extending the theory of planned behavior: Predicting the use of public transportation 1. *Journal of Applied Social Psychology*, *32*(10), 2154-2189.

Karou, S., & Hull, A. (2014). Accessibility modelling: predicting the impact of planned transport infrastructure on accessibility patterns in Edinburgh, UK. *Journal of Transport Geography*, *35*, 1-11.

Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z., 2015. Prediction Policy Problems. American Economics Review. Vol. 105, No. 5, pp. 491-95.

Menon, A. K., & Lee, Y. (2017, November). Predicting short-term public transport demand via inhomogeneous Poisson processes. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 2207-2210).

Omrani, H. (2015). Predicting travel mode of individuals by machine learning. *Transportation Research Procedia*, *10*, 840-849.

Palacio, S. M. (2018). Machine Learning Forecasts of Public Transport Demand: A comparative analysis of supervised algorithms using smart card data. *XREAP WP*, *3*.

Pereira, F. C., Rodrigues, F., & Ben-Akiva, M. (2015). Using data from the web to predict public transport arrivals under special events scenarios. *Journal of Intelligent Transportation Systems*, *19*(3), 273-288.

Team, R., et al., 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna Austria .

US Census Bureau. (2013). Appendix B. Sample Design and Weighting: 2013. Retrieved February 1, 2020, from[[https://www2.census.gov/programs-surveys/ahs/2013/2013%20AHS%20National%20Sample%20Design%20and%20Weighting.pdf?#](https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/30941/summary)](https://www2.census.gov/programs-surveys/ahs/2013/2013%20AHS%20National%20Sample%20Design%20and%20Weighting.pdf?)

US Census Bureau. (2013). Appendix D. Nonsampling and Sampling Errors. Retrieved February 1, 2020, from [https://www2.census.gov/programs-surveys/ahs/2013/2013%20AHS%20National%20Errors.pdf?#](https://www2.census.gov/programs-surveys/ahs/2013/2013%20AHS%20National%20Errors.pdf?)

US Census Bureau. (2019, June 26). AHS 2013 National Public Use File (PUF). Retrieved February 1, 2020, from[[https://www.census.gov/programs-surveys/ahs/data/2013/ahs-2013-public-use-file--puf-/ahs-2013-national-public-use-file--puf-.html](https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/30941/summary)](https://www.census.gov/programs-surveys/ahs/data/2013/ahs-2013-public-use-file--puf-/ahs-2013-national-public-use-file--puf-.html)

US Census Bureau. (2020, January 6). American Housing Survey (AHS) About. Retrieved February 1, 2020, from<https://www.census.gov/programs-surveys/ahs/about.html>

US Census Bureau. (2020, January 6). American Housing Survey (AHS) Methodology. Retrieved February 1, 2020, from[[https://www.census.gov/programs-surveys/ahs/about/methodology.html](https://www.census.gov/programs-surveys/ahs/about.html)](https://www.census.gov/programs-surveys/ahs/about/methodology.html)

US Census Bureau. (2020). AHS Codebook. Retrieved February 1, 2020, from <https://www.census.gov/data-tools/demo/codebook/ahs/ahsdict.html>

van Oort, N., Drost, M., Brands, T., & Yap, M. (2015, July). Data-driven public transport ridership prediction approach including comfort aspects. In *13th CASPT Conference, Rotterdam, The Netherlands*.

Weisbrod, G., Vary, D., & Treyz, G. (2003). Measuring Economic Costs of Urban Traffic Congestion to Business. Transportation Research Record, 1839(1), 98–106. https://doi.org/10.3141/1839-10