Predicting Station-level Ridership for Urban Rail Transit using US Census Bureau Data

Daniel Hartig

August 28, 2017

1 Introduction and Motivation

The United States is undergoing a light rail boom. New light rail systems are under construction in Los Angeles and Seattle, while extensions are being built in Denver, Minneapolis, San Francisco, Charlotte, and more. As rail transit expands in US cities, there is an opportunity to validate predictive rail ridership models.

Many machine learning models for predicting rail ridership at a station level use land-use data near stations as their feature set; otherwise, they are being made for cities outside of the United States. Models developed specifically for use with cities in the US can be used to guide future decision on which rail projects should be funded. This thesis will develop such a model.

2 Problem Description

The purpose of this research is to develop a method for predicting rail transit ridership at the individual station for a metropolitan rail network. Predictions will be based on characteristics of the area surrounding each station. We will use data from the US Census Bureau at the zip code level to estimate the number of people, jobs, and other countable characteristics within certain distances of each station. The station-specific estimates will be used as features for a regression-based machine learning model of transit ridership. Finally, after developing multiple ridership models, we will compare to determine which features and methods are most appropriate for this application.

We will use cities of varying sizes to create a general model capable of predicting rail ridership for any large cities in the US. The accuracy of each model generated will be measured on a city versus city basis. Each model will be trained on a subset of all available cities, and tested on a different subset to measure accuracy.

3 Statistical relevance

This problem is a good introduction to the challenges of machine learning applications. A first challenge is feature selection: the data set is rich in features but limited in data points. The Census Bureau produces thousands of potential products that could be used as features, but there are only a limited number of transit stations per city; for example, New York City subway, the largest in the US, has 425 stations. As a result, a model using all available features will be heavily overdetermined.

A second challenge is multicollinearity. Many of the Census products are highly correlated; for example, the total population and working age population, or total employment and total pay of all employees. Careful selection of features will be necessary to avoid high standard errors.

A third challenge is heteroscedasticity. This is important both within cities and between cities. Within cities, there is an obvious contrast in variance between a high density Central Business District and suburban areas. A good example of this is BART, San Francisco's regional heavy rail system. There are four stops in the network in the heart of San Francisco's Financial District, with population and job densities as high as anywhere in the US outside of New York City. However, much of the system runs through lower density suburbs, like the Yellow line through Contra Costa County, which have much lower absolute variance in job or population densities between stations. Between cities, we can find a contrast between the MBTA system in Boston and the DART system in Dallas. The MBTA is concentrated in the high density parts of Boston and it's immediate suburbs. Dallas has a much lower and more uniform population density than Boston or its neighbors; every city that MBTA serves has a higher population density than the city of Dallas.

4 Related Work

There is significant literature on predicting station-level ridership for rail transit networks dating back to 1980. Only in the last 10 years have machine learning methods been applied to rail transit ridership. Two important papers providing background information for this thesis are Kuby et al., (2004) and Yao (2007). These papers investigate light rail systems in nine US cities and Atlanta, respectively, and identify which factors are correlated with transit ridership using data from the 2000 US Census. These papers can be used as guides to feature selection among the much more voluminous data available today. Guerra et al, (2011) also contains important information for determining the effective radius of transit station catchments. Pulugurtha and Agurla, (2012) have important results comparing the accuracy of non-linear regression methods.

Station level modeling has been performed using distance-decay weighted regression (Gutierrez et al., 2011), support vector machines (Ostaysi et al., 2015), decision trees (Li et al., 2016), decision trees (Baek and Sohn, 2016) and more. These papers all use land use data as the features for their models. Kuby et al., (2004) and Yao (2007) both use US Census data, as is intended for this thesis. The data previously used was provided from the 2000 Census in a product called the US Census Transportation Planning Package. This product has been discontinued and was not produced for the 2010 census. One of the goals of this thesis is to recreate the data in the Transportation Planning Package and more current Census products.

Non-US cities have been the focus of machine learning approaches to transit ridership models. Cities such as Istanbul (Oztaysi et al., 2015), Seoul (Baek and Sohn, 2016), Mexico City (Duduta 2013), Singapore (Hu et al., 2016) and more have been covered in recent years. However, possibly due to a lack of good Census data, no US cities have been featured as the subject of a machine learning transit ridership model.

5 Methodology

This thesis was started as a project for STAT 672 taught by Martin Slawski. In that project, the station and network information for Chicago and Boston were collected and processed. Station level data for 13 features was then calculated and several linear regression models (simple linear regression, Support Vector Regression with three different kernels, Ridge Regression, and LASSO) applied to the feature set.

This thesis will expand on the information gathered for that project. First, we will expand the number of cities to be used as transit network sets. We will add other cities similar in size to Boston and Chicago (such as Washington DC) as well as smaller systems (potentially including Miami, Denver, Dallas, Atlanta and others). Second, we will expand the number of features in the model. The project used population, net pay for the population, number of jobs, and net pay for all jobs. This will be expanded with a variety of data from the Census Bureau including counts of populations earning more than (or less than) a certain amount of money, employment in specific categories such as universities or hospitals, population in certain age brackets, counts of rental or owner occupied properties, counts of multi-unit or single family residences, and more. A third expansion will be to add flags for interesting station properties to the model. For example, a flag for a station at an airport, for the end of a line, or for a well known tourist location such as the Mall in Washington DC or a sports stadium.

When creating the model, we will investigate alternatives to strictly linear regression modeling; specifically, we can use logarithmic transformations for some of the counting features such as population or job counts. In addition, we will investigate a variety of regularization forms to minimize the possibility of overfitting.

6 Research Plan

This thesis has three phases each with its own desired outcome.

The first objective is to develop a formal methodology for determining counting characteristics for areas near a transit station. Each station's geographic coordinates will be used with the shapefiles of nearby zip codes to determine feature counts for that station. Since in some areas rail stations are very close together, especially in the Central Business Districts of the largest cities, we need to have some way of preventing station features from being highly correlated. The solution is to draw catchments around each station so that each station is selecting data from from an exclusive area. This technique eliminates spatial dependency by making each traffic catchment independent. This formal methodology will use geographical data for the station and zip code boundaries to calculate counts for any data feature that is available on a per-zip code basis.

The second objective is to determine which available feature set has the best performance at estimating rail transit ridership. A specific challenge here is heteroscedasticity. In the previous iteration of this work, it was discovered that least squares regression presents significant modeling problems. For features such as job count within a 0.5 km radius of a transit station, the few stations in the Central Business District have much higher job counts than outlying stations. In the case of Chicago, for example, 8 of 140 stations have over 70% of the total jobs. The least squares error model minimized error by assigning a small coefficient to this feature; it simply did not predict higher ridership in the high job density areas and accepted a large error for a small number of stations.

The third objective is to provide evidence for what machine learning method is most effective at predicting transit ridership. The simple regression methods and limited feature set of the class project from which this thesis came achieved poor error scores which can be improved upon.

7 Timeline

The work on this thesis has begun already with the class project for STAT 672. The deliverable products–the thesis and the supporting code in Python–will be completed by the end of the Spring 2018 semester.

8 References

Baek, J., Sohn, K. 2016. Deep-Learning Architectures to Forecast Bus Ridership at the Stop and Stop-to-Stop Levels for Dense and Crowded Bus Networks, Applied Artificial Intelligence, 30(9): 861-885.

Duduta, N. 2013. Direct ridership model of Mexico Citys BRT and metro systems. Transportation Research Record, 2394(1): 9399.

Guerra, E., Cervero, R., and Tischler, D. 2011. The half-mile circle: Does it best represent transit station catchments? Institute of Transportation Studies, University of California, Berkeley, Working Paper UCB-ITS-VWP-2011-5.

Gutierrez, J., Cardozo, O. D., and Garcia-Palomares, J. C. 2011. Transit ridership forecasting at station level: An approach based on distance-decay weighted regression. Journal of Transport Geography, 19(6): 10811092.

Kuby, M., Barranda, A. and Upchurch, C. 2004. Factors influencing light-rail station boardings in the United States. Transportation Research Part A, 38(3): 223247.

Li, X., Liu, Y., Gao, Z., Liu, D. 2016. Decision Tree Based Station-Level Rail Transit Ridership Forecasting, Journal Of Urban Planning and Development, 142 (4).

Oztaysi, B., Yanik, S., Kahraman, C. 2015. Forecasting passenger volumes in transit systems using support vector machines: the case of Istanbul. Journal of Multiple-Valued Logic & Soft Computing, Vol. 25: 215235.

Pulugurtha, S., and Agurla, M. 2012. Assessment of Models to Estimate Bus-Stop Level Transit Ridership using Spatial Modeling Methods. Journal of Public Transportation, 15(1): 33-52.

Yao, X. 2007. Where are public transit needed – Examining potential demand for public transit for commuting trips. Computers, Environment, and Urban Systems, 31: 535-550.