A regression model for predicting rail transit ridership at the station level

Daniel Hartig

1 Introduction

The United States is undergoing a rail boom. Since 2010, new light rail lines have opened in Dallas, Los Angeles, Salt Lake City, Denver, Minneapolis, Houston, Seattle and more. A new heavy rail line opened in Washington DC, and a commuter rail system in Orlando. As transit expands in cities in the United States, there is an opportunity to validate predictive rail ridership models.

A survey of transit agencies [1] conducted by the Transit Cooperative Research Program showed that relatively few agencies are using quantitative models when forecasting ridership for new lines, extensions or stations under consideration for funding. Of the 35 agencies that responded to the survey, 29 use professional judgment and 28 use rules of thumb among one or more techniques used to generate ridership forecasts. Another method used by 22 agencies is service elasticity; a set of general transit demand response curves for changing transportation options, published by the Transportation Reserach Board [2].

For quantitative methods, the most commonly used technique–by 18 of 35 surveyed agencies–is the four-step travel demand model [3], introduced by Mannheim and Florian [4, 5]. The Mannheim-Florian model's four steps are trip generation, trip distribution, mode choice and route choice. In the trip generation phase, trip endpoints are created with as production and attraction ends. In the trip distribution step, these endpoints are paired up to generate trips; for example a residence with a a job, or hotel with a tourist attraction. In the mode choice step, trips are assigned to various transportation modes, such as personal vehicle, bus, or walking. Finally, in the route choice step, a route using that mode of transportation is chosen.

An implementation of the Mannheim-Florian model can be seen in the Seattle's Sound Transit Ridership Forecasting Methodology Report [6, 7]. The Sound Transit 3 (ST3) was a ballot measure that passed in 2016 for a $54 billion expansion of the local light rail system involving 100 km of new tracks and 37 new stations.

The ridership forecasting methodology report explained how the project's official ridership projections were developed. The regional area is divided into 785 Alternatives Analysis Zones and for each of these zones transit surveys and recorded ridership on local bus routes were used to complete the trip generation and trip distribution steps. The mode choice and route choice is done using an incremental logit model to predict changes in transit mode based on changes in transit mode availability.

Only seven of the 35 surveyed agencies used regression models to predict future transit ridership. This thesis proposes a regression-based model using data from the United States Census Bureau at the zip code level. The model will be trained on the zip code characteristics and ridership data from existing light and heavy rail transit systems and used to predict ridership on other rail transit systems.

2 Data Sources and Feature Generation

2.1 Data Sources for Predictor Variables

The zip code level data for feature generation comes from the US Census Bureau and is available at `factfinder.census.gov`. There are thousands of potential data sets available. Selection of features is guided by Kuby [8], Taylor [9], and Currie [10], which demonstrate the significance of factors such as employment, population, universities, poverty, airports, park and ride stations, and rental units. In Yao [11], a distinction is made between 'Need Index,' a series of features that depend on the characteristics around the station and are independent of the transit network, and transit network characteristics, which do depend on the transit network. To model network characteristics for each station, the sum of each characteristic for every other station within 15 and 30 minutes transit time is included as a feature of the original station. This also provides us quantitative way to express the 'centrality' dummy variable that is provided as a flag in many models [8, 12]; centrality could be proportional to the count of population or jobs within 30 minutes of a station, for example.

This model emphasizes using only features that have 'real' units. The only flag feature is for the presence of a park and ride parking spaces. Instead of using measures of land use mix as proposed in other models [12, 13], or dummy variables for universities and central business districts, the equivalent information is provided naturally as counting data in the feature set. Housing types (such as large apartment buildings versus single family homes) can stand in for land use mix, number of jobs at universities or in financial jobs can fill in for the equivalent dummy variable. A summary of the selected characteristics is provided in

Appendix A.2.

Ridership data for agencies that publish annual ridership reports is used to validate the model (see Appendix A.1). Six cities were selected for this study: Boston, Chicago, Los Angeles, Atlanta, Dallas, and Denver. Several cities were eliminated from the sample set for various reasons. A limitation of the dataset is that it does not include government employment. While state level employment is significant in all potential cities, state employment levels are relatively constant from city to city. Federal employment varies greatly, however. Washington DC was eliminated due to the large impact of un-recorded federal employment. San Francisco and Philadelphia were eliminated because they have multiple rail systems without integrated fares. New York City was eliminated because it's subway has higher ridership than all other intra-urban rail systems in the country combined.

The data closest to 2015 is used when possible to get an accurate relation between ridership and census data. The census data as well as Chicago, Dallas, and Denver's ridership statistics are from 2015. Boston's ridership is from 2014, Los Angeles' is from 2013-2014, and Atlanta's is from 2010-2013.

We translate zip code level data into transit station specific data by sampling each zip code's geographic area to determine proximity to a transit station. For each zip code near the transit network, a set of random points within that zip code is generated using rejection sampling. For each of the those points, one or more closest stations are determined. Each point is assigned to one or more station within walking distance. Counts for the characteristics of each zip code, such as population or employment, are then assigned to each station proportional to the number of points assigned to each station.

2.2 Rejection Sampling of Zip Code Shapefiles

We use a Monte Carlo method to estimate feature counts near transit stations. Sample points are generated within each zip code near the transit network. Those points are assigned to whichever stations are within walking distance of the station. The ratio of points assigned to each station to total points generated for each zip code is used to assign feature counts to each station.

The US Census Bureau provides TIGER/Line shapefiles of each zip code tabulation area (ZCTA) in the United States at `https://www.census.gov/geo/maps-data/data/tiger-line.html`. From a box drawn around the extremities of each zipcode's shape, random points are accepted if they are within the shapefile or rejected if they are outside it. Those points that are inside the shapefile are tested against author-created exclusion zones. These zones are shapes within the zip code's shapefile area that are known to not

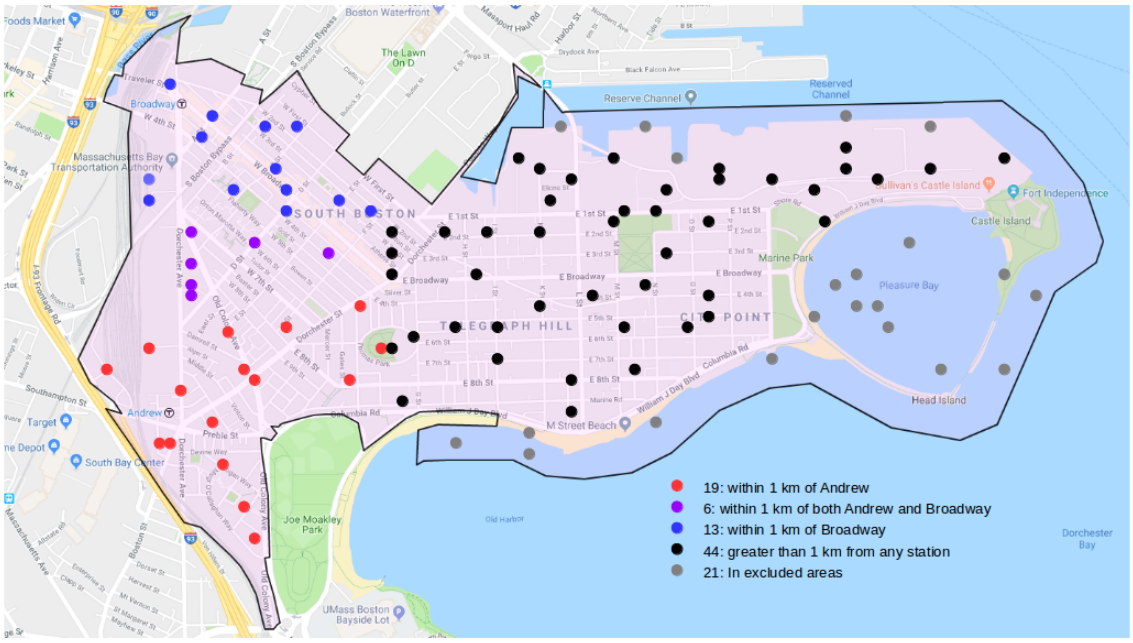


Figure 1: Illustration of nearest zip code estimation for zip code 02127.

have any population, employment, or other countable characteristics. The exclusion zones are mostly drawn over water areas, or large parks. Those points that are inside the exclusion zones are also rejected.

The remaining points are tested for their distance to any transit stations. The area within walking distance of a station is its catchment. A standard transit catchment distance for rail is one half mile (800 m), although Guerra [14] suggests that one half mile is more appropriate for population as a feature while one quarter mile (400 m) is better for employment. A case study [15] from a 2003 Montreal transit riders origin-destination survey concluded that approximately 50% of riders of the city's urban rail transit walk less than 500 meters to their stations, while 90% walk less than 1000 meters. The maximum walking distance is approximately 1500 meters. Another analysis [13] found the optimal distance for for assigning population and employment to a station was between 600 and 900 meters in straight line distance.

Given this information, we choose cutoffs of 500 meters and 1000 meters for calculating station distances. Each tested point is divided between all stations within 500 meters. If there are no stations within 500 meters, then the point is divided between all stations within 1000 meters. If no stations are within 1000 meters, that point is not assigned to any station. The total sum of points and fractional points assigned to each station is divided by the total points available. The station's portion of each of the zip code's characteristic data counts is assigned to that transit station.

An example using zip code 02127, the South Boston neighborhood of Boston, illustrates the sampling method (Figure 1). 100 random points are selected within the area of the shapefile. Of these, 21 are rejected due to exclusion areas based on water area, parks and abandoned port facilities. Of the remaining 79 points, 8 are within 500 meters of Andrew station, while 6 are within 500 meters of Broadway station. Moving out to the 1000 meter radius, 11 are within 1 km of Andrew station, for a total of 19 closest to Andrew; 7 are within 1 km of Broadway station for a total of 13 closest to Broadway; and 6 are within 1 km of both. The 6 stations within 1 km of both stations are divided between the two. The total population of South Boston is 36494. Therefore,

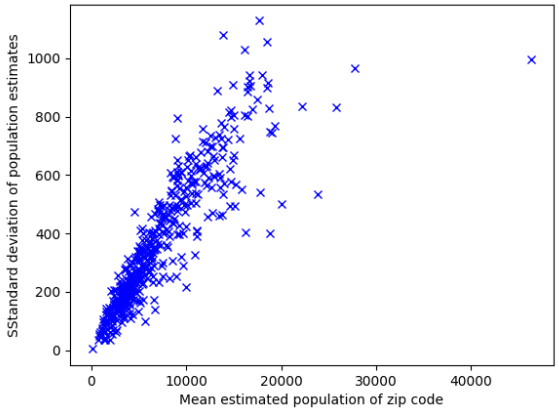
$$\frac{19 + \frac{6}{2}}{79} \cdot 36494 = 10163$$

people are assigned to Andrew station. Similarly, 7391 people are assigned to Broadway station. This calculation is performed for all countable features and all zip codes and summed total counts for each characteristic are used as a feature for each transit station.

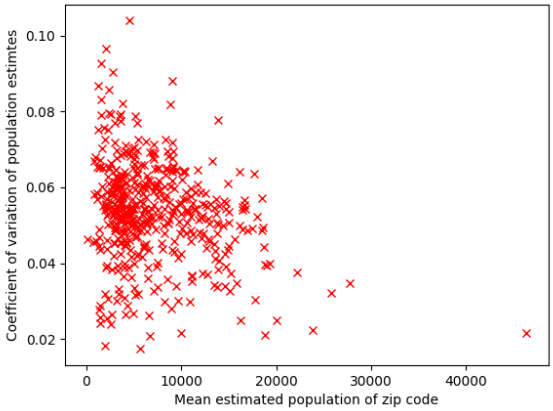
2.3 Variance of Monte Carlo estimates

With any Monte Carlo method, there is variance in the feature data generated. To keep variance to an acceptably low level, we must generate enough sample points in the Monte Carlo method. The land area of the zip codes near the studied transit networks vary in size from as small as 30 hectare in downtown Chicago to as much as 18100 hectares at the suburban end of transit line in Dallas. We to provide a balance between accuracy and processing speed, we use one point per hectare, but with a minimum limit of 1000 points per zip code. This effectively provides over 10 points per hectare in for the zip codes in the densest parts of the studied networks: the downtown areas of Chicago and Boston.

We generate 100 sets of projections for the a single feature (total population) for all stations in the six transit networks. Figure 2(a) is a graph of means against standard deviation, while Figure 2(b) is means against coefficient of variation. The standard deviation shows an expected linearly increasing relationship with the population mean. The coefficient of variation is never greater than 10.4% and generally decreases with increasing estimated mean population. The mean coefficient of variance over all studied stations is 5.3%



(a) Standard deviation versus mean by zip code



(b) Coefficient of variation versus mean by zip code

Figure 2: Indications of variance for Monte Carlo estimates of the population feature

2.4 Generation of network-dependent features

For each 'Need Index' type feature generated by rejection sampling, a set of corresponding network-based features are generated to represent the sum total of a certain characteristic (such as population or employment) within a given travel time of that station. The transit network is laid out as a graph, where nodes represent the transit stations and edges are weighted to the travel time between the stations, according to the website of the transit agency. At transfer points, there is a separate node for a single station on each line. The edge between these two nodes is the average wait time for transfer between the trains.

An illustration of the calculation of travel time between Sullivan Square and South Station in Boston is provided in Figure 3. Starting at Sullivan Square on the Orange Line southbound, there are four edge traversals totaling seven minutes to get to Downtown Crossing. From there, there is a 2.5 minute wait until a Red Line (also Southbound) train arrives, and 2.5 more minutes of travel to South Station. The total travel time is thus twelve minutes.

For each station and each 'Need Index' type feature, the total sum of all stations within 15 and 30 minutes is included as a feature of that station. These features are important for providing a measure of centrality to the network. Stations near the center of the network and at transfer points between lines will have higher counts of network features than than peripheral stations. The other important function of the network features is to provide estimates of the total scale of system ridership. The more people, jobs, and other characteristics near transit stations, the higher the overall system ridership is expected to be. This is a key component of

Figure 3: Illustration of travel time calculation for Sullivan Square to South Station, in Boston.

the model's portability between different city's transit networks.

3 Model generation and Error Metrics

3.1 Metrics for assessing accuracy of predictions

In general, the projected ridership forecasts of new transit infrastructure investments significantly overestimates transit ridership. The pioneering study in this field by Pickrell in 1989 [16] found that for ten rail projects completed between 1977 and 1985, and assessed between 1986 and 1989, the actual ridership was between 28% and 85% lower than projected. A 2006 study [17] of 25 major passenger rail projects in in 14 nations found that 21 of the projects had actual ridership below projections, with the average system ridership 48% below the projection. Accurate assessment of projection accuracy is imperative for creating ridership estimates that serve the public interest.

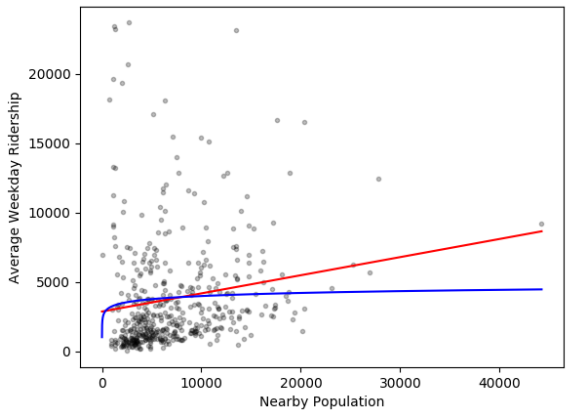
To assess the results of regression analysis, there should be two metrics used: one for the total system level ridership, and another for station level ridership. Following Pickrell, the metric for system-wide projection accuracy is standard percentage error in total system ridership, which we will refer to as system error.

$$E_{system} = \frac{\left| \sum_{i \in \text{stations}} y_{i,proj} - \sum_{i \in \text{stations}} y_{i,true} \right|}{\sum_{i \in \text{stations}} y_{i,true}}$$

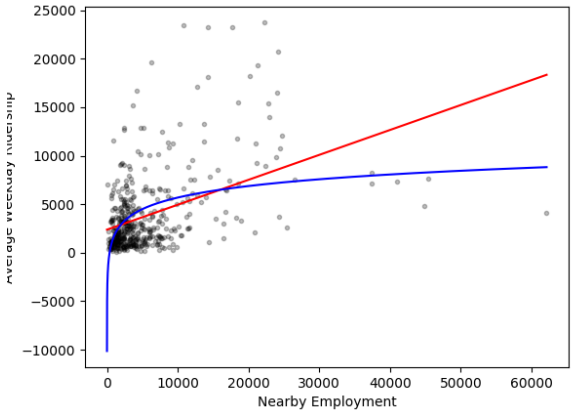
where y_{proj} is the projected ridership and y_{true} the true ridership for each station, and are summed over all stations.

Hardy [18] extends Pickrell's analysis by including absolute station error for stations on newly added sections of an existing transit network. Following Hardy, a measure of station level error on a network is summed absolute error of all station projections. The rail networks in this study vary widely in total ridership; therefore, to allow network to network comparison, this summed absolute station error can divided by total system ridership. The resulting metric for station error given a projected (y_{proj}) and actual ridership (y_{true}) is

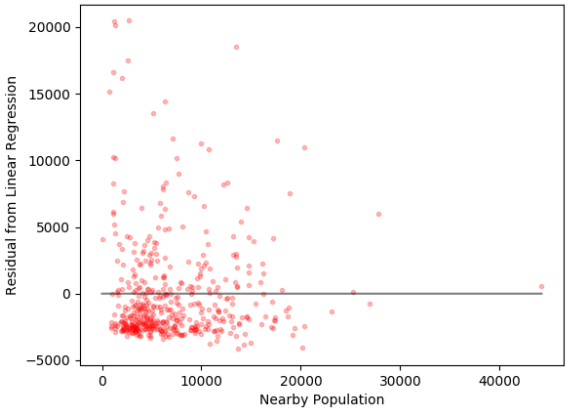
$$E_{station} = \frac{\sum_{i \in \text{stations}} \left| y_{i,proj} - y_{i,true} \right|}{\sum_{i \in \text{stations}} y_{i,true}}.$$



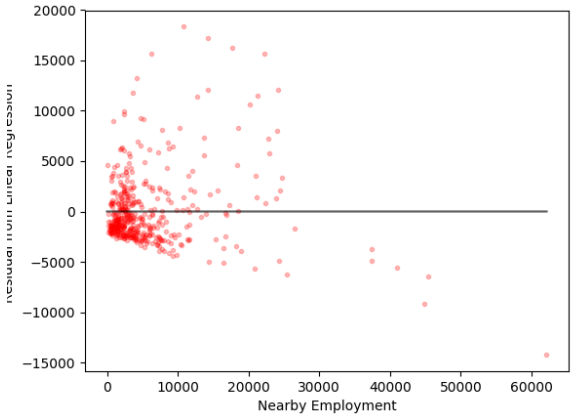
(a) Ridership against population. Linear regression in red, log regression in blue.



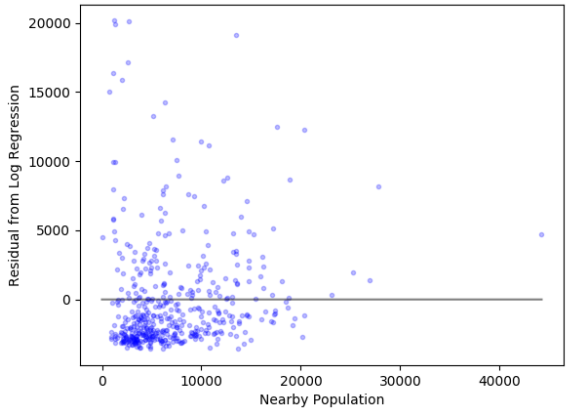
(a) Ridership against employment. Linear regression in red, log regression in blue.



(b) Residual from linear regression against population

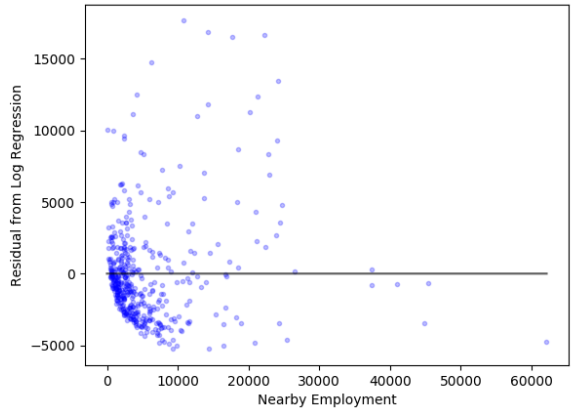


(b) Residual from linear regression against employment



(c) Residual from log regression against population

Figure 4: Analysis of population regression



(c) Residual from log regression against employment

Figure 5: Analysis of employment regression

Variable	Linear R^2	Log R^2
Population	0.0267	0.0034
Employment	0.1948	0.1878

Table 1: Regression of Population and Employment against ridership

3.2 Data distribution and regression selection

Regression models for station level ridership have used ordinary least squares regression [8, 9, 10, 12, 13] to generate predictions. We investigate the applicability of more complex regression models.

The metrics for projection accuracy depend on the absolute difference between actual and predicted ridership. This suggests that Least Absolute Deviations (LAD) regression is appropriate for this problem. Since the response variable (ridership) is counts, and the variance of ridership increases with increasing employment, we will also use Poisson regression. Finally, we will use ordinary least squares (OLS) regression as a baseline comparison, to see if the other methods have any performance advantage.

There are generally two types of feature included in this survey: those that are related to population and those that are related to employment. Features such as the population with college degrees and number of housing units will be related to total population. Features such as the number of jobs in finance or hospitality will be related to the total number of jobs. A plot of all stations' population versus ridership appears in Figure 4 with linear and logarithmic regression lines and residuals. A similar treatment for employment versus ridership appears in Figure 5. The coefficient of determination for the two regression types for both population and employment are shown in Table 1.

The domain of the dependent variable, ridership, is $[0, \infty)$, so a log link function is a reasonable assumption for this model. As demonstrated in Table 1, there does not appear to be any modeling advantage from using either an identity or logarithmic link function. We test both link functions by testing OLS and LAD with identity link, and both the identity and log link functions for Poisson regression.

There are 94 possible features, while some transit networks have as few as 38 stations. Feature selection is necessary to prevent the model from being overspecified. We use two methods of feature selection for each of the five regression types: LASSO regression and a 'brute force' method. A chart of packages used for implementation follows.

Regression	Link	package
Least Sqaures	Identity	LASSO: `glmnet` for Python
		Brute force: `statsmodels` for Python
LAD	Identity	LASSO:`flare` for R
		Brute force: `statsmodels` for Python
Poisson	Log	LASSO:`glmnet` for Python
		Brute force: `statsmodels` for Python
	Identity	LASSO: Author-created
		Brute force: `statsmodels` for Python

Table 2: Regression types and packages used in analysis

3.3 Poisson regression with Identity link

With no suitable package to perform LASSO regression using Poisson regression and the identity link, we implement this method.

For the identity link, the mean of the predicted Poisson distribution is given by

$$E(Y \mid x) = \boldsymbol{\theta}' \mathbf{x}.$$

This mean is entered in the Poisson probability mass distribution to get

$$p(y_1, ..., y_m | x_1, ..., x_m; \boldsymbol{\theta}) = \prod_{i=1}^{m} \frac{(\boldsymbol{\theta}' x_i)^{y_i} \, e^{-\boldsymbol{\theta}' x_i}}{y_i!}$$

The negative log likelihood of this distribution is the objective function

$$-\mathcal{L}(\boldsymbol{\theta} \mid X, Y) = \sum_{i=1}^{m} y_i - \log(\boldsymbol{\theta}' x_i) + \boldsymbol{\theta}' x_i.$$

This function is convex and so the minimum can be obtained using convex optimization methods.

To implement LASSO [19], we use the primal-dual interior point method from Boyd and Vandenberghe [20]. There are no equality constraints for this problem, and only one inequality constraint so we have

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_1(x) \leq 0, \quad i = 1, ..., m. \end{array}$$

Both f_0 and f_1 must be convex and twice continuously differentiable.

$$
\begin{aligned}
f_0 &= \sum_{i=1}^{m} y_i - \log\left(\boldsymbol{\theta}' x_i\right) + \boldsymbol{\theta}' x_i \\
\nabla f_0 &= \sum_{i=1}^{m} \frac{x_i \left(y_i - \boldsymbol{\theta}' x_i\right)}{\boldsymbol{\theta}' x_i} \\
\nabla^2 f_0 &= \sum_{i=1}^{m} \frac{y_i x_i^2}{\left(\boldsymbol{\theta}' x_i\right)^2}
\end{aligned}
$$

$$
\begin{aligned}
f_1 &= \sum_{i=1}^{m} |\,\theta_i\,| - t \\
\nabla f_1 &=
\end{aligned}
$$

We establish Karush-Kuhn-Tucker (KKT) conditions for optimizing the objective function subject to constraints. The primal optimization is of $\boldsymbol{\theta}^* \in \mathbb{R}^m$ so the conditions are met by

$$
\begin{aligned}
f_1 &\leq 0 \\
\lambda &\geq 0 \\
\nabla f_0 + \nabla f_1 &= 0
\end{aligned}
$$

3.4 Feature selection by LASSO regularization

Upon performing regression analysis, it becomes immediately apparent that one set of features is was different from the others. The number of students within a 15 or 30 minute transit trip is a very accurate measurement of system level ridership. These variable are represented in the model as `15net_students` and `30net_students`, respectively. We show the results of a single variable OLS regression of the one variable against ridership. The scores are the average of the six way cross-validation across the six transit networks in the study. The 'best' scoring other feature (`15net_hunits_old`; the number of housing units built before 1940 within a 15 minute transit ride) is shown for comparison.

The system error scores for `30net_students` and `15net_students` are much lower than for any other variable, while the station error for these features are also lower than any other features. As we will see, the single-feature OLS of either of these features produces a model that is approximately as good as any other model we will develop. This raises questions about the relationship between the feature and the response variable. It is possible that the population of students within walking distance of a transit station is driven by the availability of local transit, and not the other way around. In that case, number of students is not a

Variable	System Error	Station Error
30net_students	0.1016	0.5961
15net_students	0.0946	0.6197
15net_hunits_old	0.2854	0.6700

Table 3: Error for single variable OLS for selected features

Regression Type	Result	Boston	Chicago	Los Angeles	Atlanta	Dallas	Denver	Average
OLS	System	0.3711	0.6120	0.1499	0.4218	0.5187	0.2491	0.3871
	Station	0.6444	0.8290	0.5077	0.5736	0.8829	0.7553	0.6988
	# Features	2	7	17	10	10	10	9.3
Poisson	System	0.4094	0.8617	0.0543	0.4867	0.4859	0.4416	0.4566
	Station	0.6470	1.0397	0.5863	0.5701	0.9137	0.9021	0.7765
	# Features	5	13	26	12	27	26	18.2
LAD	System	0.6161	0.8431	0.4455	0.4340	0.3347	0.2889	0.4937
	Station	0.6519	1.0023	0.6824	0.5809	0.7502	0.7560	0.7373
	# Features	36	52	48	54	47	53	48.3

Table 4: OLS, Poisson, and LAD LASSO regression results

valid explanatory variable. Since the relationship is unclear, but the features are outliers, we will remove all features derived from number of students from the model.

OLS and Poisson LASSO regression are performed using the `glmnet` package in R. LAD LASSO is performed using the `flare` package. The `glmnet` does automatic λ selection; for `flare` the author created a comparable λ search algorithm. Table 4 shows the results for the three regression types.

There is a large variance in number of features chosen, even within a single regression type. The number of features chosen varies significantly both between transit networks and between regression types. The large number of features selected by LAD LASSO regression may result from the different λ selection methods used by the computing packages. The average of 48 features selected in this method is far too high. The transit networks have between 37 and 138 stations. Several of the LASSO selections have too many features, and some of the poor accuracy may be attributable to overfitting.

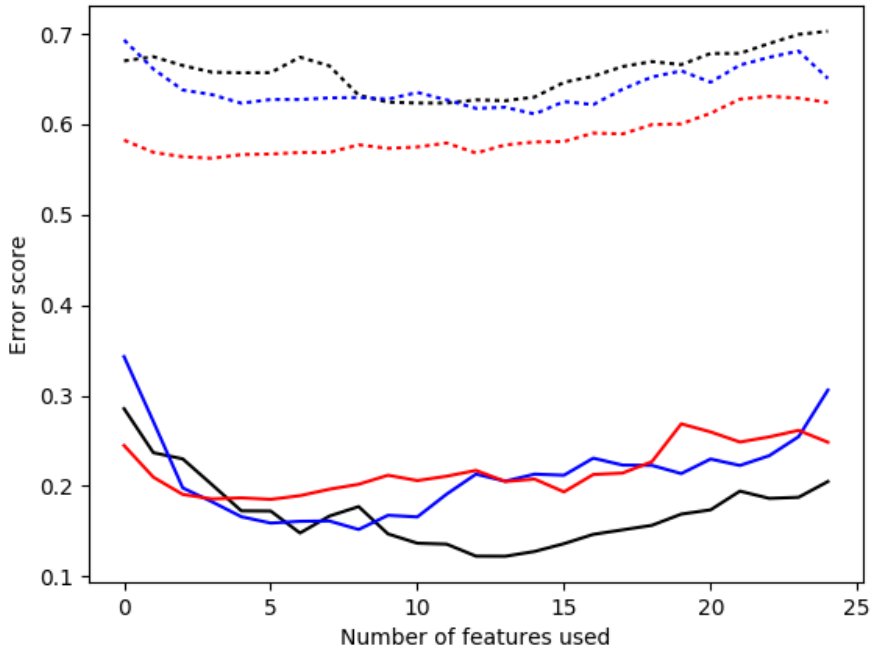


Figure 6: Graph of System (solid line) and Station (dotted line) score against number of features for OLS (black), Poisson (blue), and LAD (red) regression

3.5 Feature selection by brute force

Since the data set for this problem is small–with only 466 total stations–we can validate the LASSO results using a 'brute force' approach. The brute force method is a greedy search of all possible features to find the best 'path' to a regression solution. First, all features are checked by regression against ridership. Since we have two 'score' metrics, we use the average of system and station error to determine the best feature in each step. Each feature is checked in six-way cross validation. Each of the six cities is used as the test set while the other five are used for the training set. The average over all six cross-validation runs for each metric is what is reported in the charts below.

After choosing the one feature that yields the highest score, we then select a second feature to add to the regression in in the same way, and iteratively add more features. The subsequent steps are multiple regressions using all of the already chosen features. We select the fist 25 features with this method and graph the resulting system and station error scores in Figure 6.

In general, each of the error scores decreases with the addition of new variables up to a point, and then increases again. The minimum for system and station errors do not coincide with each other for any of the

Regression Type	Min System Err	Min Station Err	# Features Selected	Avg System Err	Avg Station Err
OLS	0.1222	0.6234	10-17	0.1342	0.6317
Poisson	0.1518	0.6113	4-16	0.1806	0.6255
LAD	0.1851	0.5623	2-16	0.1998	0.5718
OLS LASSO	0.1499	0.5077	9.3	0.3871	0.6988
Poisson LASSO	0.0543	0.5701	18.2	0.4566	.07765
LAD LASSO	0.2889	0.5809	48.3	0.4937	0.7373

Table 5: Results of brute force regression analysis, with comparison to LASSO results

three regression methods, and there is a range of features which produce very similar error scores. The three regression methods each perform roughly as well as each other Error scores are averaged across the six-way cross validation. The minimum scores from each regression type are recorded in Table 5 along with the average score over the range of good features.

The brute force regression significantly outperforms the LASSO regression in creating feature combinations that can accurately predict ridership in unknown transit networks.

A summary of selected features is given in Appendix B. For the brute force results, the best one feature is selected by six way cross validation for each step, from one to twenty five. Those features that fall with the range of good estimates, as shown in Table 5, are marked in the Appendix.

For the LASSO feature selection, each of the six cross-validated models produces a unique set of features. Therefore, each feature is selected up to six times. We wish to avoid selection of so many features as to cause overfitting. Based on the error scores from the brute force approach, ten to fifteen features seems to be the optimal choice for all regression types. To get this number of features for each LASSO regression type, those features selected by at least four of six cross validated runs are marked in the Appendix. For the LAD LASSO, the number of features selected by LASSO is unusually high, so those features selected by at least five of size cross validated runs are selected, to avoid saturated models.

4 Conclusion and Future Work

The Mannheim-Florian four step model fundamentally depends on existing transit ridership information. For example, the starting point of a four step analysis for a new light rail line would be an existing bus line that runs a similar, hopefully identical, path. The advantage of this regression model is that it creates a new estimate from different sources, independent of any knowledge of the current system.

All three regression types are able to give systemwide ridership estimates with 20% of the true value for identified sets of between 10 and 15 features. This does not depend on any previous ridership measurement of the transit system in question and compares favorably with the 48% average ridership prediction error reported for 25 transportation networks in Flyvbjerg [17]. It is justifiable to use this regression based ridership model to produce and validate system-level estimates for new construction intra-urban rail transit in the United States.

Future work in on this model could proceed in two directions. The first direction is to continue improvement of the source data. This project used only zip code shapes to estimate counting features, since job and housing data was available only for the zip code, but the population features exist in more granular detail at the Census Tract level. The author generated exclusion zones which were designed to prevent job and people from being located in parks and water could be improved by incorporating detailed city land use maps. Finally, there is a major deficiency in source data is that government workers are not included in the data sources used by the model. Of particular concern are university employment; university jobs within walking distance was a selected feature in five of the six models. Several large universities are located on the transit networks of this study and were not accounted for, such as Illinois-Chicago, Colorado-Denver, and UMass-Boston. For other cities that have very large public universities, like Minneapolis or Columbus, this would significantly affect the validity of any estimates.

The second direction is improvements of the model itself. The feature summary for this paper only analyzed selection of a feature by either the LASSO or brute force method. There remains to be done an analysis of the magnitude and direction of each feature's coefficient, to ensure that frequently selected features are significant. The R package `glmnet` is capable of performing ElasticNet regression, but for this work only ℓ_1, LASSO regularization was used. For OLS and Poisson, a mixed regularization may be able to improve the performance of the feature selection.

A Data sources

A.1 Ridership data

Los Angeles:	http://libraryarchives.metro.net/DPGTL/Ridership/RailActivityByStationFY2014.xls
Chicago:	http://www.transitchicago.com/assets/1/ridership_reports/2015_Annual.pdf
Atlanta:	http://documents.atlantaregional.com/transportation/TFB_2014_v17.pdf
Boston:	http://archives.lib.state.ma.us/bitstream/handle/2452/266319/ocm18709282-2014.pdf
Denver:	http://www.rtd-denver.com/documents/serviced/lrt-activity-08-2015.pdf and
	http://www.rtd-denver.com/documents/serviced/lrt-activity-Jan-April-2016.pdf
Dallas:	https://www.dart.org/about/dartreferencebookmar16.pdf

A.2 US Census feature data sources

All feature data is accessed through the American Factfinder website at factfinder.census.gov.

Population	Table DP05, Item HC01_VC03
Population, 18 and under	Table DP05, Item HC01_VC03 - Item HC01_VC32
Population, 65 and over	Table DP05, Item HC01_VC37
Housholds	Table S1101, Item HC01_EST_VC02
Households with Children	Table S1101, Item HC01_EST_VC06
Families	Table S1101, Item HC01_EST_VC010
Population with at least Bachelors degree	Table S1701, Item HC01_EST_VC34
Population in labor force	Table S1701, Item HC01_EST_VC37
Employed population	Table S1701, Item HC01_EST_VC38
Full-time employed population	Table S1701, Item HC01_EST_VC47
Population living at greater than 500% of poverty level	Table S1701, Item HC01_EST_VC56
Population living at less than 200% of poverty level	Table S1701, Item HC01_EST_VC01 - HC01_EST_VC59
Housing units	Table DP04, Item HC01_VC03
Single-family detached housing units	Table DP04, Item HC01_VC14
Housing units in duplexes or townhouses	Table DP04, Items HC01_VC15 + HC01_VC16
Housing units in structures of 3-9	Table DP04, Item HC01_VC17 + HC01_VC18
Housing units in structures of 10+	Table DP04, Item HC01_VC19 + HC01_VC20
Housing units built before 1940	Table DP04, Item HC01_VC36
Housing units built after 2000	Table DP04, Item HC01_VC27 + HC01_VC28 + HC01_VC29
Housing units occupied by owner	Table DP04, Item HC01_VC65
Housing units occupied by renter	Table DP66
Number of Jobs	Table CB1500CZ11, Item EMP
Total pay of all jobs	Table CB1500CZ11, Item PAYANN
Number of jobs at hospitals	Table CB1500CZ21, NAICS code 622, Estimated
Number of jobs at universities	Table CB1500CZ21, NAICS code 6113, Estimated
Number of jobs in hospitalitiy field	Table CB1500CZ21, NAICS code 72, Estimated
Number of jobs in finance field	Table CB1500CZ21, NAICS code 52, Estimated
Number of jobs in professional fields	Table CB1500CZ21, NAICS code 54, Estimated
Number of jobs in entertainment fields	Table CB1500CZ21, NAICS code 71, Estimated

B Summary of selected features

		OLS		Poisson		LAD	
Variable Name	Description	LASSO	BF	LASSO	BF	LASSO	BF
30net_medical	Medical jobs within 30 min	x	x	x	x	x	
near_hospitality	Hospitality jobs within walking	x		x	x	x	x
near_university	University jobs within walking	x	x	x	x		x
parking	Flag for available parking	x		x	x	x	x
15net_medical	Medical jobs within 15 min		x		x	x	x
near_business	Business jobs within walking		x		x	x	x
near_entertainment	Entertainment jobs within walking		x	x	x	x	
15net_hunits_attached	2-4 unit housing within 15 min	x		x		x	
15net_hunits_medium	5-19 unit housing within 15 min	x		x		x	
15net_hunits_old	Housing build before 1940 within 15 min		x		x	x	
30net_entertainment	Entertainment jobs within 30 min		x		x		x
30net_hunits_large	20+ unit housing within 30 min		x		x	x	
near_employment	Total employment within walking		x	x		x	
near_medical	Medical jobs within walking		x		x		x
near_pop_old	Population over 65 within walking	x		x	x		
15net_employed	Employed population within 15 min					x	x
15net_hospitality	Hospitality jobs within 15 min	x				x	
15net_household	Housholds within 15 mins					x	x
15net_university	University jobs within 15 minutes		x		x		
near_emp_pay	Total employee pay within walking		x			x	
near_finance	Finance jobs within walking		x				x
near_house_w_child	Households with children within walking		x		x		
near_hunits_detached	Single unit housing within walking				x	x	
near_hunits_owner	Owner occupied houses within walking					x	x
near_hunits_renter	Renter occupied houses within walking		x		x		
15net_bachelors	Population with degree within 15 min					x	
15net_emp_pay	Total employee pay within 15 minutes					x	
15net_entertainment	Entertainment jobs within 15 minutes						x
15net_hunits_renter	Renter occupied houses within 15 min			x			
15net_pop_old	Population over 65 within 15 minutes	x					
30net_hospitality	Hospitality jobs within 30 min			x			
30net_hunits_medium	5-19 unit housing within 30 min					x	
30net_hunits_vacant	Vacant housing units within 30 min					x	
30net_population	Total population within 30 min						x
near_emp_full_time	Population employed full time within walking						x
near_hunits_attached	2-4 unit housing within walking						x
near_hunits_large	20+ unit housing within walking		x				
near_hunits_new	Housing built after 2000 within walking						x
near_labor_force	Population in labor force within walkin					x	
near_pop_rich	Population over 500% of pov. level within walking					x	
near_population	Total population within walking		x				

References

[1] Daniel Boyle. *Fixed-Route Transit Ridership Forcasting and Service Planning Methods*. Transportation Reserach Board, Washington, DC, 2006.

[2] Transportation Research Board and National Academies of Sciences, Engineering, and Medicine. *Traveler Response to Transportation System Changes Handbook, Third Edition: Chapter 1, Introduction*. The National Academies Press, Washington, DC, 2013.

[3] Michael G. McNally. *The Four-Step Model*, chapter 3, pages 35–53. 2008.

[4] Marvin L Manheim. *Fundamentals of transportation systems analysis*. Cambridge, Mass. : MIT Press, 1979. Includes index.

[5] Michael Florian, Marc Gaudry, and Christian Lardinois. A two-dimensional framework for the understanding of transportation planning models. *Transportation Research Part B: Methodological*, 22(6):411–419, December 1988.

[6] Sound Transit. ST3 Regional High-Capacity Transity System Plan: Transit Ridership Forecasting Methodology Report. Technical report, March 2015.

[7] Sound Transit. ST3 Regional High-Capacity Transity System Plan: Addendum to Transit Ridership Forecasting Methodology Report. Technical report, April 2015.

[8] Michael Kuby, Anthony Barranda, and Christopher Upchurch. Factors influencing light-rail station boardings in the united states. *Transportation Research, Part A: Policy and Practice*, 38(3):223–247, 3 2004.

[9] Brian D. Taylor, Douglas Miller, Hiroyuki Iseki, and Camille Fink. Nature and/or nurture? analyzing the determinants of transit ridership across us urbanized areas. 2008.

[10] Graham Currie, A Ahern, and Alexa Delbosc. Exploring the drivers of light rail ridership: An empirical route level analysis of selected australian, north american and european systems. 38:545–560, 05 2011.

[11] Xiaobai Yao. Where are public transit needed - examining potential demand for public transit for commuting trips. *Computers, Environment and Urban Systems*, 31:535–550, 2007.

[12] Matthew Durning and Craig Townsend. Direct ridership model of rail rapid transit systems in canada. 2537:96–102, 01 2015.

[13] Javier Gutirrez, Osvaldo Daniel Cardozo, and Juan Carlos Garca-Palomares. Transit ridership forecasting at station level: an approach based on distance-decay weighted regression. *Journal of Transport Geography*, 19(6):1081 – 1092, 2011. Special section on Alternative Travel futures.

[14] Erick Guerra, Robert Cervero, and Daniel Tischler. Half-mile circle. *Transportation Research Record: Journal of the Transportation Research Board*, 2276:101–109, 2012.

[15] Ahmed El-Geneidy, Michael Grimsrud, Wasfi Rania, Paul Ttreault, and Julien Surprenant-Legault. New evidence on walking distances to transit stops: Identifying redundancies and gaps using variable service areas. 41, 01 2014.

[16] Don Pickrell. Urban rail transit projects: Forecast versus actual ridership and costs. final report. 10 1989.

[17] Bent Flyvbjerg, Mette Skamris, and Sren L. Buhl. Inaccuracy in traffic forecasts. 26:1–24, 01 2006.

[18] Matthew H. Hardy, Soongwan Doh, Junyang Yuan, Xin Zhou, and Kenneth J. Button. The accuracy of transit system ridership forecasts and capital cost estimates. *International Journal of Transport Economics / Rivista internazionale di economia dei trasporti*, 37(2):155–168, 2010.

[19] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[20] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.