Predicting Transit Demand on the Chicago 'L' Train

August Posch
Northeastern University
posch.au@northeastern.edu

Abstract

Time series for Chicago 'L' train daily ridership by station were analyzed, and ridership prediction models were compared for three types of stations. Clustering and mapping of stations revealed that stations with certain attributes tend to be located in specific geographic areas. Supervised machine learning on each of three representative stations revealed that Extra Trees models are the best class of model for daily demand prediction tasks as a function of the last two weeks or five weeks of daily demand. These results clear the way to operationalize a real-time prediction system, which would help the agency relieve bottlenecks and help riders find less crowded vehicles.

Introduction

This study develops a novel way to predict transit demand, helping transit operators run the system more efficiently and helping riders have a more pleasant riding experience.

Urban theorists describe the city in ways both sociological and geographic, with Bettencourt defining the city as "an agglomeration of social links" (2013). He says everything about a city can be determined from how easy it is for people to connect with each other, from economic flourishing to cultural output to health of residents. Bettencourt's view inspires me to improve cities by increasing people's opportunities for social links, achieved via public transit.

The United States has a history of car-oriented development, which comes with many downsides. Car emissions contribute to climate change; road noise, engine noise, and horns are unpleasant; car crashes endanger pedestrians; and highways and parking lots inefficiently take up public space. In summary, US cities have not been human centered. But they could be. The United States is just beginning to catch up to Europe and Asia as it builds out transit systems, makes them efficient, and encourages people to ride them. The prime challenges are convenience, comfort, and speed to compete with cars.

This project addresses the comfort disparity between transit and cars, by helping transit riders avoid crowded vehicles. Before the Covid-19 pandemic, and increasingly post-pandemic, crowding was a major factor prompting people to avoid public transit. However, there may be a data-oriented solution to crowding. Noursalehi, Koutsopoulos, Zhao (2021) are developing a crowding prediction platform that can alert riders at a station to wait for the next, less-crowded subway train, and this information can also help the transit agency make tactical decisions about where to send vehicles to relieve bottlenecks. The current project develops a daily demand (ridership) prediction model by station for the Chicago 'L' train. These predictions could help riders avoid crowding, and they could serve as an input to improve the vehicle-level prediction model mentioned above.

Background (Literature Review)

Types of Transit Data

Ge et al. reviewed types of transit data in their recent paper (2021).

Automatic Vehicle Location (AVL) data shows where all vehicles in a transit system are located at any one time. The GTFS-R format is a data standard used by Google Maps as well as app developers, allowing riders to view an app to see where the next vehicle is and when it is expected to arrive at their stop. AVL data is widely available, but does not contain information about how many riders are on each vehicle, so it is not appropriate for demand prediction.

Automated Fare Collection (AFC) data tracks ridership by station. Transit agencies typically issue a card, e.g. CharlieCard (Boston), MetroCard (New York), or ORCA (Seattle). Riders tap the card when

boarding transit; in some systems, riders also tap the card as they disembark. AFC data is widely used within transit agencies and serves as a rich data source, perfect for demand prediction. The main obstacle to using this data is convincing the agency to share it – usually agencies keep this data close to the vest, as sharing it could expose them to political attacks and privacy concerns.

Automated Passenger Counting (APC) data tracks ridership by vehicle. Machine vision and depth sensors count the number of people boarding and alighting each vehicle. This could be an excellent data source for demand prediction, as long as the accuracy of the machine vision is known. However, this technology only exists in some transit systems, and not widely in the United States.

Crowding Prediction

Both before pandemic and increasingly in 2022, crowding was an issue dissuading people from riding transit. Less-crowded vehicles lead to happier riders, which leads to more people choosing transit over cars. Noursalehi, Koutsopoulos, Zhao (2021) used time series of past demand at 15-minute intervals to predict demand in the future, e.g. in 15 minutes or in 30 minutes. They developed a model that predicts crowding on individual vehicles, based on the number of people who entered stations earlier on the vehicle's journey, as well as vehicle location information. They also propose an information-sharing platform inside stations so that riders can confidently choose whether to board the current vehicle or wait for the next vehicle, which may be much less crowded.

Our study aims to make the best predictions of the number of riders beginning their trip at a station. This could serve as an input to improve the vehicle-level crowding model and the information shared with the riders, ultimately improving their experience.

Chicago 'L' Train

The 'L' is Chicago's high-volume rail rapid transit system, akin to what other cities call a metro or a subway. In Chicago it became known as the 'L', or 'El', due to significant portions of its track being elevated. It currently consists of 145 stations, although only 139 stations are considered in our study. The Chicago 'L' is America's second-busiest by average weekday ridership (measured pre-Covid), behind only the New York City Subway.

Out dataset is provided by the Chicago Transit Authority (CTA) on the open Chicago Data Portal. It contains rides by day by station from 2001 to present (Chicago Data Portal 2022). Here, a "ride" is defined as a station entry, so one is counted each time a person passes through a turnstile to enter a station. We can think of this as a type of AFC data. In the same open portal, the CTA provides a stations dataset which includes the geographic location of each station, various ID numbers, and description of which 'L' lines use each station; this dataset allows us to clean the rides data and plot results on a map (Chicago Data Portal 2018). We deeply appreciate the folks at the CTA, as theirs was the highest-resolution publicly available transit ridership dataset.

Methods

Computing Tools

We used Python and associated data science tools: Numpy and Pandas for data manipulation; Geopandas for geospatial data manipulation and map visualizations; Contextily for basemaps; Matplotlib for visualizations; Statsmodels and Scikit-learn for machine learning.

Data exploration and cleaning

We stated with a stations dataset and a rides dataset. We explored the data by making time-series plots of daily ridership of a station in a year. We prepared the data for further analysis by filtering out some stations and combining others. Stations that were established after start of 2015 or discontinued before end of 2019 were removed from the dataset. Other stations required fixing ID numbers, which could be figured out using the rides time series and the supplemental ID numbers in the stations dataset. Finally, it was ambiguous how to classify certain stations, as some stations have an underground subway station that is connected indoors to an aboveground elevated station. In these cases a judgement call was made, informed by how rides at stations had been tracked by the CTA.

Understanding types of stations

Summary statistics were calculated for daily rides for each station using the 2015-2017 timeframe. This timeframe was chosen because it is completely before the Covid-19 pandemic, and we are interested in creating a model that is useful to riders in a non-pandemic situation. The summary statistics were lag 1

autocorrelation, mean, variance, standard deviation, and scaled standard deviation (a heuristic defined as standard deviation divided by mean). Summary statistics were stored in the stations geodataframe and plotted on a map to get a sense of the geospatial distribution of attributes.

We wanted to see what types of stations existed and ensure our prediction models would work across the major types of stations. Therefore, we performed k-means clustering with k=3, based on lag 1 autocorrelation, mean, and scaled standard deviation, all of which were pre-scaled using StandardScaler. Cluster labels were plotted in the space of these three attributes as well as on a geospatial map. Finally, we found the most representative station for each cluster, by starting with cluster center and searching over all stations for which is closest (Euclidean distance) in the space of these variables. The representatives were marked in the aforementioned cluster label plots.

Demand prediction

Using each of the three representative stations chosen above, and using a 2015-2019 timeframe, an array was created in a format suitable for machine learning. Each array was organized so that each day's number of rides is predicted by the preceding 35 days. That is, each row looks like this: the target y is the number of rides on a particular day; the input x is 35 numbers, each of which is the number of rides on a preceding day ("rides 35 days ago", "rides 34 days ago", "rides 33 days ago", … , "rides 2 days ago", "rides yesterday").

Each array was split into train/validation/test datasets in a 0.6/0.2/0.2 ratio of the rows. The same random seed was used for splitting each array, so that one station's training set corresponded to the same dates as another station's training set

The goal of machine learning was to predict demand (rides) at each station based on a time series of preceding days at the same station. A variety of machine learning models were tried, including various model classes (ARIMA, linear regression, random forest, extra trees, and support vector regression) and used two different amounts of input data (14 preceding days of rides, versus 35 preceding days of rides). For all models, we trained on the training set and evaluated using root mean square error (RMSE) on the validation set. In addition, we performed 5-fold cross-validation on the training set and evaluated using mean RMSE and standard deviation of RMSE among the five folds.

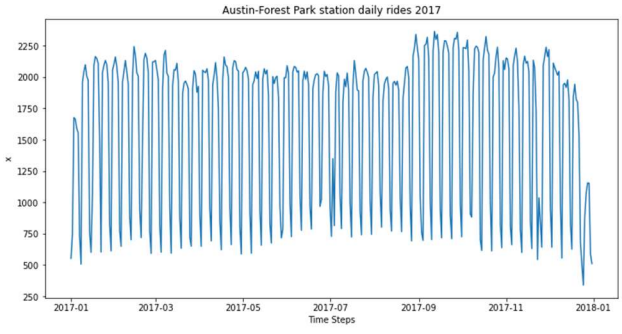
Hyperparameter tuning was not extensive; we mostly used the Scikit-learn defaults. These are the only departures from Scikit-learn defaults: n_estimators = 500 in the random forest and the extra trees; C = standard deviation of the training values and epsilon = 0.2 * standard deviation of the training set in the support vector regression. The ARIMA model was the only one not provided by Scikit-learn, and in order to use it we generated each validation set prediction solely based on the preceding days in the validation set series – so the training dataset was not used. We used hyperparameters order = (1,0,0), seasonal_order = (1,0,0,7). Given the author's inexperience with this API, it is possible this ARIMA methodology is flawed.

The best models were chosen in a two-part process. First, for each of the three stations the best model was chosen based on RMSE on the validation set, and the model was then re-used on the test set, and the test set RMSE values were reported. Second, cross-validation results were analyzed in the context of station attributes.

Results

Exploratory Data Analysis

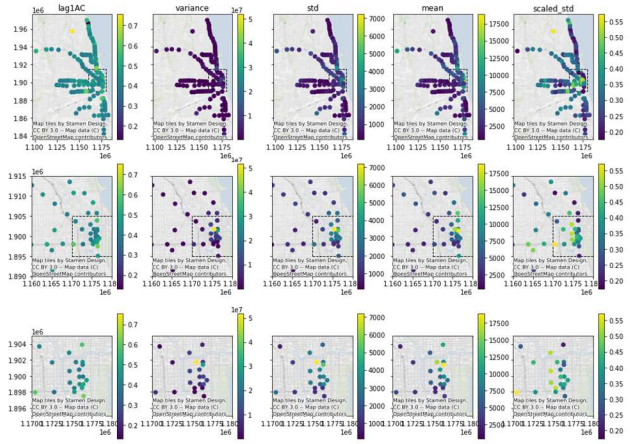
The exploratory data analysis revealed that for most stations, weekdays have much higher ridership than weekends, creating a 7-day cycle in ridership. Most stations also have a slight seasonal change, most apparent in summer versus autumn, and an extreme dip during the week of Christmas. While stations have a lot of these same characteristics, the magnitudes of the 7-day cycles and the seasonal changes may be different across stations. See below (Austin-Forest Park station daily rides 2017) for one example. (See Appendix A1 for larger plot.)

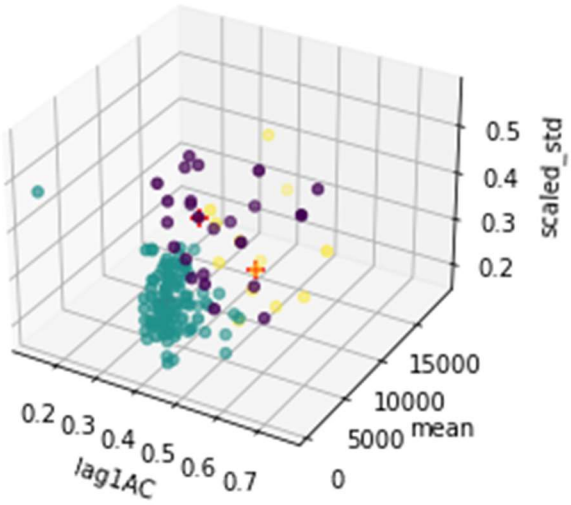


Understanding types of stations

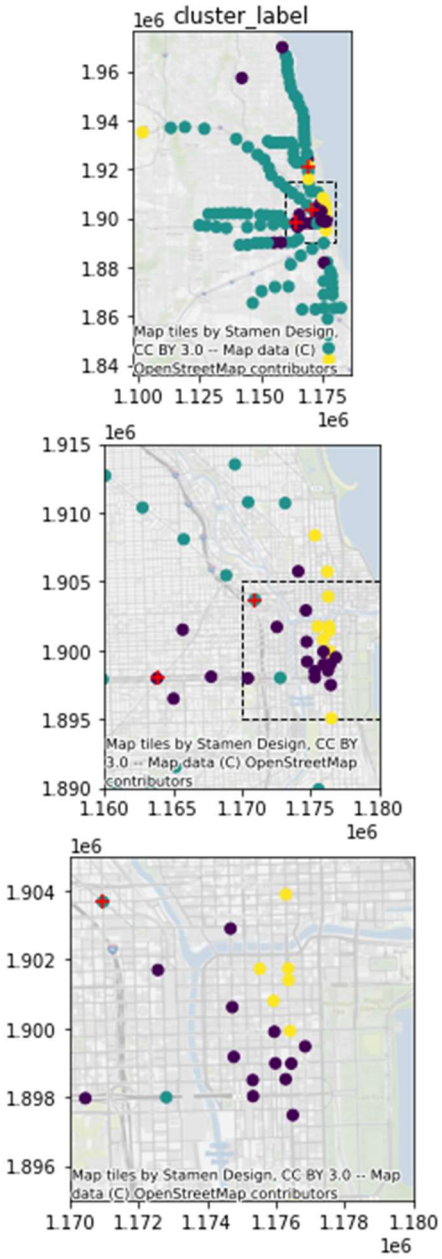
Maps of summary statistics revealed that lag 1 autocorrelation, mean, and scaled standard deviation

show the most visible differences across space. Some outliers are apparent (for example, a northern station with very high lag 1 autocorrelation – research revealed it suffered closure for many months due to embankment collapse), but there were few enough outliers that we did not tailor our analysis to them, because we want to make demand prediction models that are most useful to people using normal stations. See below for maps – each column shows a particular attribute with station locations mapped onto Chicago at three different magnifications. (See Appendix A3 for larger map.)





K-means clustering and mapping of the cluster labels revealed distinct characteristic of each cluster. The purple cluster has high lag 1 autocorrelation, high scaled standard deviation, and low mean, and tends to be located in the city center. The aqua cluster has low lag 1 autocorrelation, low mean, low scaled standard deviation, and tends to be located in the outer part of the city. The yellow cluster has high mean and low scaled standard deviation, and tends to be located in the city center. Below is a 3D plot of clusters labels in the statistic-space, as well as a Chicago map at three magnifications showing cluster labels at station locations. (See Appendix A4 and A5 for larger versions.)



The representative stations are Illinois Medical District (purple), Grand (aqua), and Belmont (yellow). In the plots above these representatives are marked with a red plus (+) symbol.

Demand prediction

For context of the scale of these RMSE numbers, note the following mean ridership values: Illinois Medical district at 2452.5 rides, Grand at 2360.8 rides, Belmont at 11549.3 rides. (See Appendix A6 for full context, the statistics of the stations.)

Supervised machine learning revealed that Extra Trees models always performed the best, as measured by RMSE on the validation set. For Illinois Medical District, the 14-day Extra Trees model performed the best, and it had RMSE of 394.3 on the test set. For Grand, the 35-day Extra Trees model performed the best, and it had RMSE of 271.7 on the test set. For Belmont, the 35-day Extra Trees model performed the best, and it had RMSE of 1259.3 on the test set.

See below for model comparison on the validation set. (See Appendix A7 for larger versions.)

RMSE on validation set

		Illinois Medical District	Grand	Belmont
35 day models	ARIMA	1059.4	700.6	2555.0
	Linear regression	528.8	342.2	1877.1
	Random forest	483.9	300.2	1775.6
	Extra Trees	485.8	**292.5**	**1733.9**
	SVR	508.8	322.3	1841.1
14 day models	ARIMA	1052.0	697.6	2518.3
	Linear regression	548.6	364.2	2019.9
	Random Forest	499.1	311.4	1740.3
	Extra Trees	**483.4**	302.9	1800.7
	SVR	507.0	315.4	1822.3

These results show that Extra Trees tends to be the best model, and that usually the next best model is Random Forest, followed by Support Vector Regression, followed by Linear Regression. All of these other models were relatively close behind Extra Trees in performance.

The ARIMA model always had the worst performance. Each time, the RMSE was nearly the same as RMSE from predicting the mean of the input series. Thus, ARIMA using the methodology in this study did not improve upon a naïve prediction. It is unclear whether this study's methodology was flawed for ARIMA, or

whether ARIMA is actually the wrong model for this situation. Because of this, we left out ARIMA from the rest of the analysis.

Cross-validation revealed not only that Extra Trees models perform the best, but also how reliably they can be expected to win out over other models. The cross-validation showed that for all three stations, 14-day Extra Trees models had the lowest mean RMSE. For Illinois Medical District, the mean RMSE was 354.9, with a standard deviation of 48.0, and the next-best model class had mean RMSE of 368.8 – about a quarter of a standard deviation away. For Grand, the mean RMSE was 265.0, with a standard deviation of 44.1, and the next-best model class had mean RMSE of 282.5 – about a third of a standard deviation away. For Belmont, the mean RMSE was 1688.3, with a standard deviation of 571.3, and the next-best model class had mean RMSE of 1727.2 – about a fifteenth of a standard deviation away.

See below for cross-validation results. (See Appendix A8 for larger version.)

mean RMSE among 5 folds of training set

standard deviation of RMSE among 5 folds of training set

		Illinois Medical District	Grand	Belmont
35 day models	ARIMA			
	Linear regression	407.2	294.9	1941.9
		33.4	40.6	532.4
	Random forest	368.8	286.0	1727.2
		38.5	39.7	600.6
	Extra Trees	**354.9**	**265.0**	**1688.3**
		48.0	44.1	571.3
	SVR	383.4	282.5	1738.7
		33.1	45.2	566.1
14 day models	ARIMA			
	Linear regression	447.9	325.1	1985.5
		41.7	45.5	**513.7**
	Random Forest	383.6	293.3	1809.8
		54.8	46.0	549.5
	Extra Trees	368.9	271.7	1701.8
		53.8	**38.8**	539.0
	SVR	375.1	282.9	1731.0
		41.4	41.6	570.7

For all models tested in cross validation, the mean RMSE was an amount between 11% and 18% of the mean daily ridership of the station – meaning that, on average, our model predictions are not too far off for the purpose of predicting crowding. (See Appendix A9 for complete results in this percentage format.)

Cross-validation also showed that for each of the three stations, no model had a substantially different standard deviation of RMSE, so all these error estimates are similarly reliable.

Discussion

To reduce emissions and create a human-centered living experience for the increasing populations living in American cities, we need to make transit better so that people choose transit over cars. The results of this study solve one piece of how to make transit better: reducing crowding. By predicting transit demand, we can inform riders and allow them to choose less-crowded vehicles, and we can inform transit agencies to allow them to make tactical decisions that further relieve crowding. Together, these improvements will make people's transit rides less crowded and more comfortable, encouraging them to continue to choose transit.

Although various types of stations exist, we see that for all station types, we can expect an Extra Trees model to produce the best day-level predictions. However, based on cross-validation analysis and statistics, the advantage of the Extra Trees model is rather small for Illinois Medical District and Grand, and exceedingly small for Belmont. Belmont's defining attribute is its high mean daily ridership. Belmont and stations like it serve downtown entertainment and workplaces, or act as transfer points to other modes of transportation. Our results suggest that such busy stations are harder to predict. Perhaps this is due to one-time events such as baseball games or concerts.

With some extra tuning, the models in this study are a solid approach for predicting ridership at the day level. Here is the thought process as to why that is the case. Imagine we look at the cross-validation results and assume RMSE is Normally distributed among folds, and assume we've decided on a 35-day Extra Trees model type. For Illinois Medical District, most of these models have error in the range of 12-17% of mean daily ridership; for Grand, most of these models have error in the range of 9%-14% of mean daily ridership; for Belmont, most of these models have error in the range of 9%-20% of mean daily ridership. With further parameter tuning and on-line model training, we expect to be able to improve the models as well, so let's figure we can usually count on a model that has RMSE around 10% of mean daily ridership. Even then, the squaring aspect of RMSE accentuates the worst errors, meaning that most of the time, our prediction for a given day has error under 10% of the day's ridership. Thus, most days, we could predict this station's number of rides within 10%. We think this kind of accuracy is useful for crowding prediction.

Predicting demand at the daily level by station is only one piece of the puzzle. To get a fuller picture of crowding in a transit system, we would like to be able to look at intra-day data and integrate it with information about the location of each vehicle, in order to determine which actual vehicles are crowded. This study's methodology could be extended to intraday demand prediction at, for example, 15-minute intervals. This kind of time resolution would allow riders and agencies to make tactical choices throughout the day, and such predictions would nicely feed into a vehicle-level crowding prediction platform like the one described by Noursalehi, Koutsopoulos, and Zhao. In fact, this author is doing a research project this summer utilizing Seattle light rail intraday demand data.

This study's supervised machine learning models were trained only using preceding ridership values. Future models could be trained using additional data, like weather, and engineered features such as day of week and month of year, as well as preceding ridership values from other stations geospatially nearby.

This study's models were evaluated using root mean square error, which averages all the prediction errors while weighting egregious errors more heavily. A future study should evaluate models primarily on upper-tail errors. Upper-tail error is the difference between transit vehicles being overcrowded or merely somewhat crowded; this distinction is much more important to the rider than the difference between empty and somewhat crowded, which was a highly-weighted error in this study.

It may even be beneficial to view this as a classification problem, rather than a regression problem. Based on a particular station, its frequency of service, and the capacity of its trains, we could determine a threshold of ridership at which we consider it overcrowded. Then, we are simply trying to classify each new chunk of time as "Yes, overcrowded" or "No, not overcrowded". This

approach could be extended to include a three classes, such as “Uncrowded”, “Crowded But Not Full”, and “Full”. Providing a classification as output also is more legible to the rider, so classification lets us skip a translation step when we’re broadcasting model outputs to an app.

We envision a future in which everyone has access to crowding predictions at their fingertips. Given a real-time data stream, we could set up a website with a model from this study that tells riders whether to expect a crowded vehicle or not. We think this tool would create a better rider experience and help agencies address problems in real time. Here’s to a future of human-centered cities with excellent public transit.

References

Bettencourt, Luis, 2013. “The Kind of Problem a City Is.”

Chicago Data Portal. 2018. “CTA - 'L' (Rail) Stations – Shapefile”

Chicago Data Portal. 2022. “CTA - Ridership - 'L' Station Entries - Daily Totals”

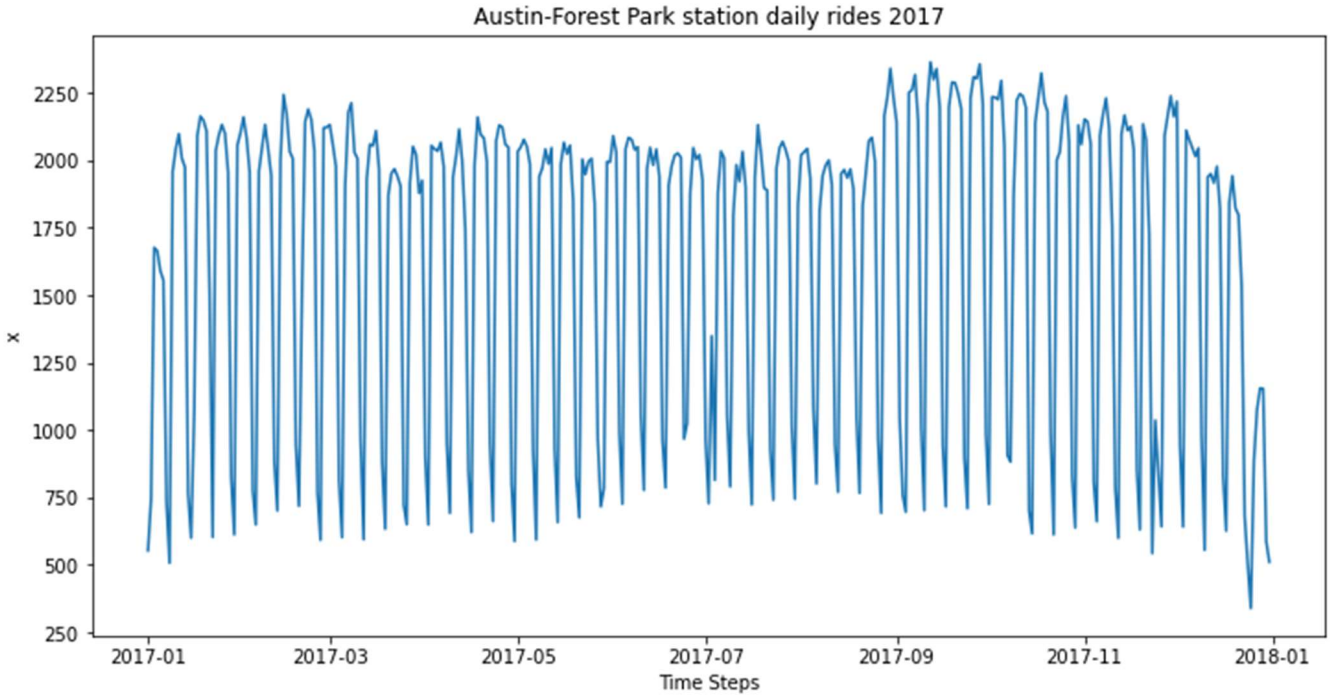
Ge, Liping et al. 2021. “Review of Transit Data Sources: Potentials, Challenges and Complementarity.”

Noursalehi, Peyman, Haris Koutsopoulos, and Jinhua Zhao. 2021. “Predictive decision support platform and its application in crowding prediction and passenger information generation.”

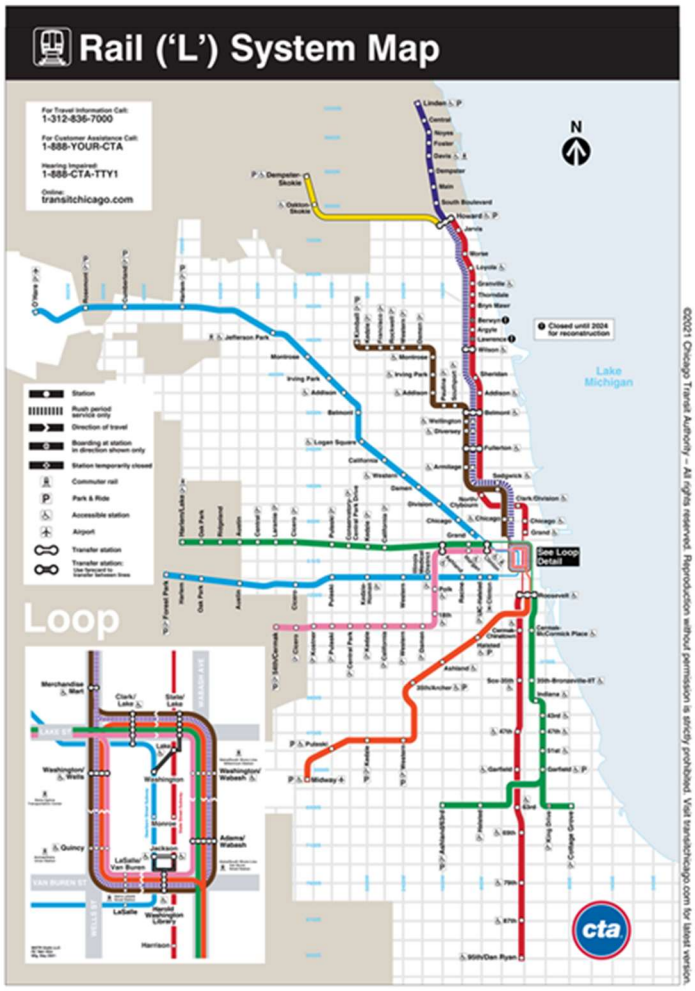
Appendix

Here are all the figures referenced in the text, in a larger size.

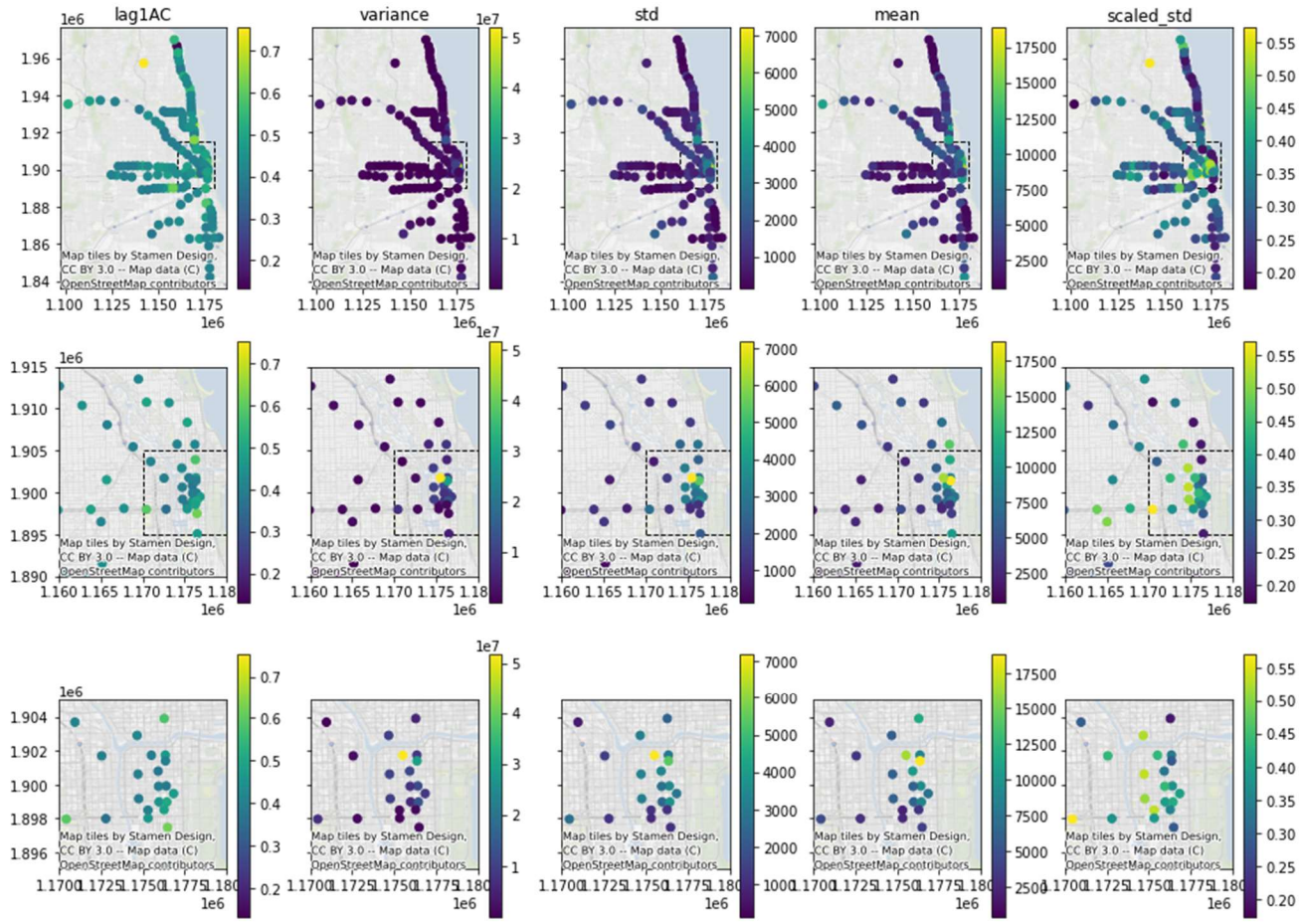
A1. Exploratory data analysis for Austin-Forest Park station, 2017.



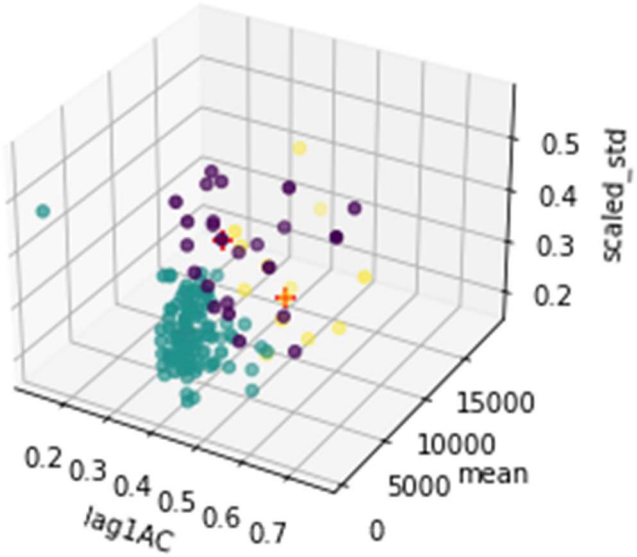
A2. Chicago transit map by the CTA.



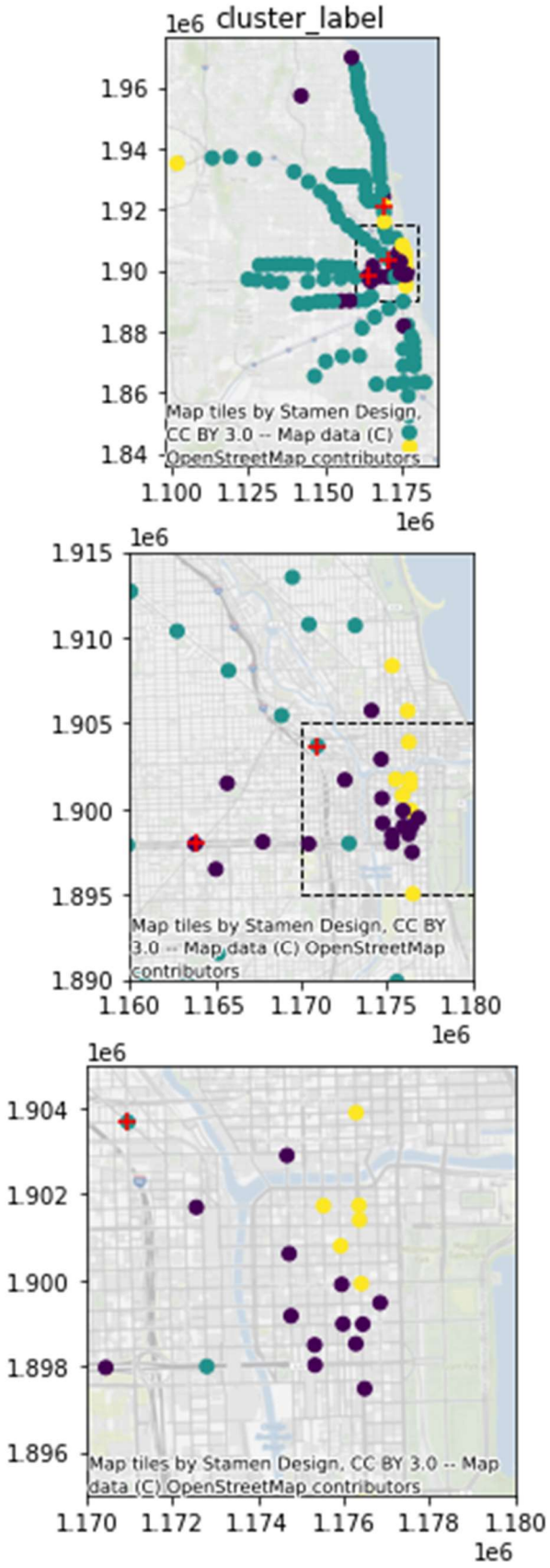
A3. Summary statistics mapped at three magnifications.



A4. Cluster labels on a 3D plot in the statistics-space.



A5. Cluster labels mapped in Chicago locations.


cluster_label
1e6
1.96
1.94
1.92
1.90
1.88
1.86
1.84
1.100
1.125
1.150
1.175
1e6
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors
1e6
1.915
1.910
1.905
1.900
1.895
1.890
1.160
1.165
1.170
1.175
1.180
1e6
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors
1e6
1.904
1.902
1.900
1.898
1.896
1.170
1.172
1.174
1.176
1.178
1.180
1e6
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors

A6. Statistical attributes of this study's stations of interest.

statistic		Illinois Medical District	Grand	Belmont
	Cluster color	purple	aqua	yellow
	Lookalikes located	downtown	outer	downtown
	Mean daily rides	2452.5	2360.8	11549.3
	Scaled Std	0.4843	0.2938	0.2284
	Lag 1 Autocorr	0.4923	0.4117	0.4553

A7. Supervised machine learning train-val-test results, shown in two different table configurations.

RMSE on validation set										
	35 day models					14 day models				
	ARIMA	Linear regression	Random forest	Extra Trees	SVR	ARIMA	Linear regression	Random Forest	Extra Trees	SVR
Illinois Medical District	1059.4	528.8	483.9	485.8	508.8	1052.0	548.6	499.1	**483.4**	507.0
Grand	700.6	342.2	300.2	**292.5**	322.3	697.6	364.2	311.4	302.9	315.4
Belmont	2555.0	1877.1	1775.6	**1733.9**	1841.1	2518.3	2019.9	1740.3	1800.7	1822.3

RMSE on validation set				
		Illinois Medical District	Grand	Belmont
35 day models	ARIMA	1059.4	700.6	2555.0
	Linear regression	528.8	342.2	1877.1
	Random forest	483.9	300.2	1775.6
	Extra Trees	485.8	**292.5**	**1733.9**
	SVR	508.8	322.3	1841.1
14 day models	ARIMA	1052.0	697.6	2518.3
	Linear regression	548.6	364.2	2019.9
	Random Forest	499.1	311.4	1740.3
	Extra Trees	**483.4**	302.9	1800.7
	SVR	507.0	315.4	1822.3

A8. Supervised machine learning cross-validation results.

mean RMSE among 5 folds of training set				
standard deviation of RMSE among 5 folds of training set				
		Illinois Medical District	Grand	Belmont
35 day models	ARIMA			
	Linear regression	407.2	294.9	1941.9
		33.4	40.6	532.4
	Random forest	368.8	286.0	1727.2
		38.5	39.7	600.6
	Extra Trees	**354.9**	**265.0**	**1688.3**
		48.0	44.1	571.3
	SVR	383.4	282.5	1738.7
		33.1	45.2	566.1
14 day models	ARIMA			
	Linear regression	447.9	325.1	1985.5
		41.7	45.5	**513.7**
	Random Forest	383.6	293.3	1809.8
		54.8	46.0	549.5
	Extra Trees	368.9	271.7	1701.8
		53.8	**38.8**	539.0
	SVR	375.1	282.9	1731.0
		41.4	41.6	570.7

A9. Supervised machine learning cross-validation results, as a percentage of mean daily rides.

mean RMSE among 5 folds of training set (as a percentage of mean ridership)		These are heuristic scores to try to compare model performance across stations.		
standard deviation of RMSE among 5 folds of training set (as a percentage of mean ridership)				
		Illinois Medical District	Grand	Belmont
35 day models	ARIMA			
	Linear regression	16.6%	12.5%	16.8%
		1.4%	1.7%	4.6%
	Random forest	15.0%	12.1%	15.0%
		1.6%	1.7%	5.2%
	Extra Trees	**14.5%**	**11.2%**	**14.6%**
		2.0%	1.9%	4.9%
	SVR	15.6%	12.0%	15.1%
		1.3%	1.9%	4.9%
14 day models	ARIMA			
	Linear regression	18.3%	13.8%	17.2%
		1.7%	1.9%	**4.4%**
	Random Forest	15.6%	12.4%	15.7%
		2.2%	1.9%	4.8%
	Extra Trees	15.0%	11.5%	14.7%
		2.2%	**1.6%**	4.7%
	SVR	15.3%	12.0%	15.0%
		1.7%	1.8%	4.9%