Fare is Fair: Creating Chicago transit zones for ’L’ travel

Nathaniel Smith | smithnj@bu.edu | BU CAS CS 504 | Spring 2019

I. INTRODUCTION

The City of Chicago currently has a flat-rate fare of $2.50 for entry into the Elevated Rail (“L") transit system. Other metropolitan transit systems such as Transport for London's Tube network have fare zones where, depending on the entry and exit into the system, a fare may be higher or lower. While the "L" infrastructure is currently incapable of tracking when passengers exit the system, could Chicago benefit from zone-based fares?

To determine if a zone-based system for the 'L' would benefit Chicago, three goals are presented:

1. How are the zones determined?
2. What would the zones look like?
3. Are the metrics used statistically significant?

Using various public Chicago datasets, fare zones can be created for the "L" network that encourage transit in burdened areas of the network while offsetting this fare deficit by charging riders in more stressed parts of the network a higher fare.

The datasets obtained are as follows:

1. ’L’ Station Ridership - Station Entry Totals (2012-2018) *[Chicago Data Portal]*
2. CTA ‘L’ (Rail) Lines *Chicago Data Portal, Converted to .geojson, Uploaded to datamechanics.io]*
3. Chicago Neighborhood Socioeconomic Indicators (US Census - 2010) *[Data.gov]*
4. Chicago Taxi Trip Data (2018, Limit 3.5M entries grabbed.) *[Data.gov]*
5. ’L’ Station Location Data *[Chicago Data Portal, Converted to .geojson, Uploaded to datamechanics.io]*
6. Census Tract Boundaries Location Data *[Chicago Data Portal]*

This project creates the following datasets:

1. **ctapopularity**: Aggregates CTA ridership totals per station from Dataset 1 and calculates per-station variance from the mean.
2. **taxiagg**: Aggregates taxi demand per Chicago Census Community Area Number (“CAN”) from Dataset 4. “Demand Metric” is the total number of rides beginning and/or ending in a CAN.
3. **stationhardship**: Assigns CAN and Census Hardship Index to each Station ID obtained from Datasets 3 and 5
4. **metrics**: Pairs each station with the proper hardship index (from “stationhardship”), popularity (from “ctapopularity”), and taxi demand (from “taxiagg”).
5. **zones**: After running the k-means algorithm for zone assignment, pairs stations from metrics with their assigned zones.
6. **kmeans.data**: metrics data is scaled before k-means analysis. This scaled data is stored in this database for visualizations.
7. **kmeans.centers**: K-means cluster centers are stored in this database for visualization purposes.

II. METHODS

Data Retrieval & Selection

Data was gathered and selected from the respective data portals. Datasets obtained from either Data.gov or the Chicago Data Portal were retrieved using the SODA API via the SodaPy library. Queries made used specific restrictions such as a certain timeframe or a restricted number of data entries.

Other datasets were retrieved and manually converted to the geoJSON filetype. These datasets were then uploaded to datamechanics.io.

All data was then sent into its respective Pandas DataFrame, converted to the JSON filetype, and finally inserted into a MongoDB database. Any project files tasked with retrieval of data follow the template “get_DATA.py” and any tasked with transformations (such as projection, selection, and aggregation) follow the template “create_DATA.py”.

To calculate the station popularity metric, monthly ridership totals per station between 2012 and 2018 were retrieved and aggregated to achieve a total sum of 2012-2018 ridership per station. The new dataset “ctapopularity” contains each station’s CTA ID, ridership total, and variance of ridership total from the mean ridership.

The taxi demand and socioeconomic burden metrics as retrieved do not come with an assigned “L” station. These two statistics are based on Chicago Community Area Numbers. Thus, before any station could be assigned these two metrics, a CAN was assigned to each station. However, 22 CTA “L” stations do not reside in Chicago proper. Because of this, any station residing outside of official city boundaries will not have these two metrics. These stations are dropped using the Pandas “dropna” method before any further analysis was performed.

The taxi demand metric originally pulls from 3.5M taxi ride entries available from the Chicago Data Portal. As there are over 122M+ entries, retrieval of all rides during the 2018 year would take an extremely large amount of time to download, parse, and store in MongoDB. As such, the arbitrary 3.5M entries proved to be an acceptable amount to continue with the project. Data was then aggregated to calculate the total number of pick-ups and drop-offs in all CAN’s throughout Chicago. The new dataset “taxiagg” contains an entry for each Community Area Number and its respective Demand Metric.

The City of Chicago uses its own statistical data on the city to assign each Community Area a “hardship index” taking into account socioeconomic burden indicators such as the percent of households living below the federal poverty level. Each station, already being matched with a CAN, is then linked to the hardship index of the surrounding neighborhood. This new dataset “stationhardship” contains a Station ID, CAN, and that CAN’s hardship index.

Once these three metrics were gathered, a final dataset “metrics” was generated with each entry being a single “L” station containing the Station ID, Popularity, Hardship Index, and Taxi Demand.

K-Means Analysis & Zone Creation

The metrics dataset was then processed through the k-means algorithm available in the scikit-learn cluster library. All metrics were scaled before k-means processing with the scikit-learn preprocessing library. Metrics were scaled as each individual metric is an arbitrary type of measurement and it is recommended to scale data when working with such an environment.

K-means analysis was performed four times with eight, six, four, and three clusters. Each cluster center was saved to the dataset “kmeans.centers” and scaled data was saved to the dataset “kmeans.data.” After each run of the algorithm, each station’s assigned cluster was saved.

For the purposes of creating fare zones, each k-means cluster will represent an individual zone, since it is assumed each cluster will have similar metric values that make it suitable for a fare zone.

Statistical Analysis

Correlation coefficients were determined for the three metrics and their relation to each other. These correlation coefficients are between Popularity and Hardship, Hardship and Taxi Demand, and Popularity and Taxi Demand. The SciPy statistics library was utilized to perform this analysis.

Visualizations

Each run of the k-means algorithm and the cluster/zones it created were visualized using the Plotly library. Each visualization shows all stations and the cluster/zone center.

To create map visualizations of the newly created fare zones, the Folium python library was utilized. Data was read from the “zones” dataset and each station’s zone determined its marker color. To enhance maps, the CTA “L” line map was inserted showing the path between stations. Interactive map visualizations were created for the 8-cluster/zone and 4-

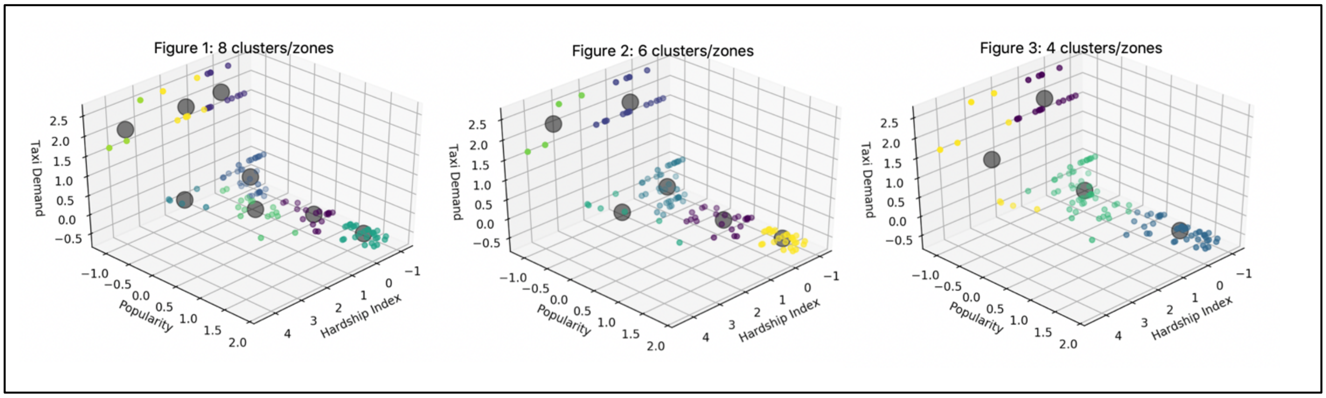

Figure 1: 8 clusters/zones

Figure 2: 6 clusters/zones

Figure 3: 4 clusters/zones

cluster/zone runs of the k-means algorithm. For the purpose of visualization, creating maps of the 6-cluster/zone and 3-cluster/zone data was deemed non-essential and thus not performed.

III. RESULTS & DISCUSSION

Statistical Analysis

Correlation coefficient calculations show the following results:

Metric Relation	Coefficient
Popularity and Hardship	0.50736
Popularity and Taxi Demand	0.519861
Hardship and Taxi Demand	-0.61339

These coefficients show somewhat intuitive and expected results. In areas where there is large station activity, such as the downtown "Loop" area, there is also a relatively large amount of taxi demand. In socioeconomically burdened areas there is negative correlation with taxi demand. This is largely expected as these areas likely prefer and rely on lower-cost transit such as the "L" or CTA bus.

A significant indication on the benefit of a new fare-zone system is the correlation between station popularity and hardship. In general, stations that are in more burdened areas see increased ridership. These citizens actively rely on the "L" more so than less burdened areas.

Metric Visualization

Visualization of the metrics, including their assigned zones, show notable conclusions for the "L" network and, in a larger scope, the city of Chicago itself. Stations are either in areas that show high taxi demand or are not. When scaled, no station has a taxi demand in the 0.5-1.5 range.

The metric visualizations also confirm there are no irregularities in cluster centers determined by the k-means algorithm. However, seen in the difference between clusters/zones in Figure 2 and 3, transitioning from 6 zones to 4 zones does group stations together that have large metric variance, especially in regard to taxi demand. If the CTA were to implement fare zones, this factor should possibly be taken into account.

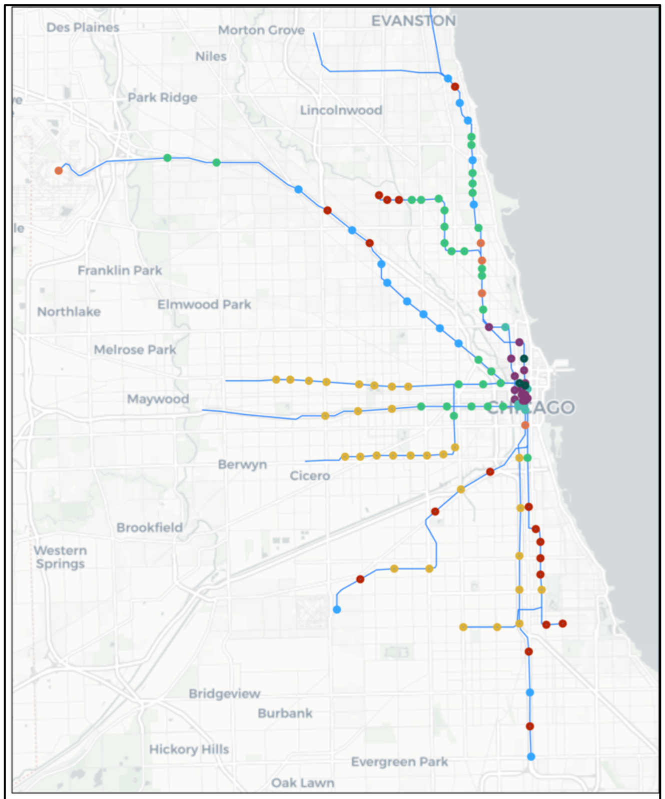

Figure 4: Generated "L" map network from k-means clustering. K-means (k=8) clustering performed on all stations in Chicago proper with three metrics.

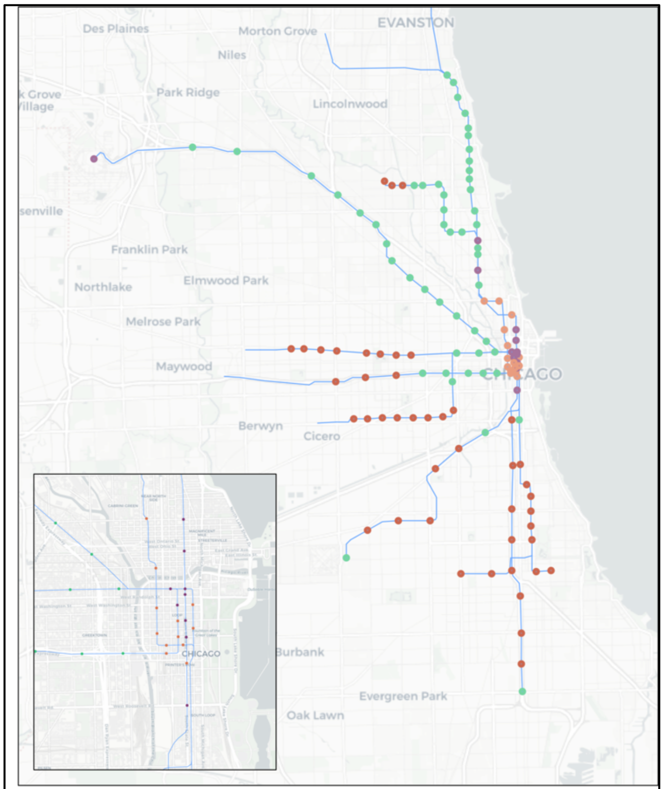

Figure 5: Alternate Generated "L" map network from k-means clustering. K-means (k=4) clustering performed on all stations in Chicago proper with three metrics.

Map Visualizations
Utilizing k-means analysis on station metrics proves to be useful for generating theoretical "fare zones" for the CTA "L" network. In both map visualizations, a strong disparity between stations in the downtown "Loop" area and the more distant neighborhoods is shown.

Transit systems such as London's Tube similarly have a more expensive central zone in the busiest part of the city. An unexpected result of including the hardship index metric is seeing a similar divide between downtown stations and those further out in the city.

Stations in the South of Chicago are in their own zone separate from stations in the North due to the severe socioeconomic disparities in the city. Therefore, Southern stations should see a lower fare than Northern stations.

Basing fares off of these zones would promote transit in disadvantaged areas while possibly increasing revenue due to a likely higher fare in the "Loop". However, in zone maps generated from larger cluster amounts, the CTA would have to take practicality into mind. Figure 4, where k=8, CTA lines regularly switch between different fare zones. This would likely confuse customers and be generally unpopular. In Figure 5, where k=4, lines more or less maintain a singular fare zone outside of the downtown area and, as a result, this map may be more practical.

Overall, the CTA could see benefits from a zone-based approach to fares but would have to take practicality into account before deciding on a final map.

IV. FUTURE WORK

This project does not touch on assigning actual fares to the newly created zones. One route the project could take in the future would be to assign these fares and explore their impact. An important consideration is how much revenue the CTA currently brings in as a result of the "L" and how much new zones would change that total revenue. It's possible the change would drastically alter the amount of revenue generated and arguably this is an even more critical factor than the fare zones themselves.