Metro and Bike-sharing Systems in Washington DC, Boston and San Francisco: a data-driven approach

Lily Liu
Cornell University
School of Operations Research and Information Engineering

May 28, 2020

Contents

1 Introduction

With the rapid popularity of bike-sharing systems, stationed and now dockless bikes have played an increasingly important role in people's daily travel routine. More and more people use shared bikes to commute from home to nearby metro stations; others may even choose to ride bikes instead of walking, taking a metro or taxi.

The high level goal of this project is to understand how each method in the transportation system interacts with each other. In particular, we focus on the metro system and bike-sharing system. Since the bike-sharing systems complete the trip chain by connecting metro stations with points of interest, it is necessary to understand the various trans-shipment characteristics of bike-sharing systems for metro stations. Through case studies in Washington DC (DC), Boston and San Francisco (SF), we get a better understanding of operating characteristics of dockless and stationed bikes near metro stations and compare the ridership patterns between metros and shared bikes.

In this project, we use a data-driven approach. We collected data from dockless bikes system in DC through a live API and also worked on various published datasets for metro ridership and stationed bikes trip history. We prepossessed and analyzed data through various methods, including K-means clustering.

2 Operating Characteristics of Dockless Bike-Sharing Systems near Metro Stations: Case Study in DC

2.1 Background

There are 91 metro stations in the greater DC area, spanning 6 lines and 177 miles of route. According to District Department of Transportation, there are currently two companies, HelBiz and Jump, operating dockless electric bicycles. This case study aims at analyzing the operating characteristics of a dockless bike-sharing system near metro stations using data from Jump and the locations of 40 metro stations in around the DC city center. "Operating characteristics" here refers to the temporal usage of bikes near metro stations.

Field	Content
Name	Columbia Heights
Address	3030 14TH STREET NW
Latitude	38.927846
Longitude	-77.032554

Table 1: Structure of metro stations data

2.2 Data collection and pre-processing

For this case study, metro stations data and Jump bikes location data were collected and analyzed.

2.2.1 Metro stations data

We obtained metro stations information (including name and address) from Open Data DC. We then used the Washington Metropolitan Area Transit Authority API to obtain the corresponding latitude and longitude of each station. After merging and pre-processing, the final data contain information including name, ID, address, latitude and longitude, as shown in Table 1.

2.2.2 Jump bike locations data

We obtained the Jump bike locations data by running Python script every two minutes through a live API. The data was collected from June 2, 2019 to June 24, 2019. The raw data has the structure as shown in Table 2. There are 1,548,086 rows in the dataset. The relevant information includes time, bike id, latitude and longitude.

2.2.3 Data pre-processing

Since we are interested in the dockless bikes near metro stations, before analyzing data, we first combined two previous datasets and calculated the distance (based on latitudes and longitudes) between each pair of metro station and dockless bike. The structure of the pre-processed data is shown in Table 3.

Field	Content
bike_id	bike_173131
is_disabled	0
is_reserved	0
battery	79
lat	38.90584
lon	-76.9866
name	XJJ803
time	2019-06-07T10:26:02
epoch	1559917562
last_updated	2019-06-07T10:25:33

Table 2: Structure of Jump bike locations data

Field	Content
Time	2019-06-07T10:26:02
Metro station name	Columbia Heights
Metro latitude	38.927846
Metro longitude	-77.032554
Bike ID	bike173131
Bike latitude	38.90584
Bike longitude	-76.9866
Distance (feet)	297.27

Table 3: Structure of integrated data

Metro station ID	0:00	0:15	0:30	...	23:45
1	0.94	1.12	1.06	...	0.22
2	9.94	11.56	10.19	...	11.38
...	...	...	...	...	...
30	1.06	0.71	0.47	...	0.24

Table 4: Extracted features

2.3 Cluster analysis

2.3.1 Feature extraction

After data pre-processing, we first filtered only data entries such that the origin or destination of a Jump bike is a metro station. Due to the flexibility of dockless bikes, people usually take and return bikes near metro stations when they use the metro as a connecting transportation method. We set the threshold to be 300 feet; bikes within 300 feet of a metro station are considered as "near metro stations". Out of 40 stations, there are 30 of them that has nearby bike activities.

For every metro station, the number of nearby Jump bikes was counted and averaged for every time period for weekdays and weekends, respectively. We separated weekdays and weekends because we expected different travel patterns. The feature extracted data structure is shown in Table 4. For example, the value "0.94" in the cell with ID of Metro Station = 1 means that, during the data collection period, there was on average of 0.94 bikes within 300 feet of the metro station No. 1 during 0:00-0:15.

2.3.2 Cluster analysis

We applied k-means clustering to analyze the specific activity patterns. In particular, we want to partition those 30 stations (with nearby dockless bike activities) identified above into k clusters such that stations within a cluster has similar activity patterns. Before applying k-means clustering, we first separated data into two groups: weekdays and weekends since we expected very different activity patterns during weekdays (when people mostly commute to work) and weekends (when people use public transportation for other purposes than commuting).

It is critically to choose the value of k: on one hand, as k increases, the partitioning error monotonically decreases and the classification performance

increases; on the other hand, the classification becomes meaningless when k becomes extremely large. We used the Elbow Method to choose the value of k at the inflection point such that once the k value exceeds the point, the improvement in the partitioning error decreases significantly. As shown in Figure 1(a), we chose k value for weekdays to be 5 and similarly, we chose k value for weekends to be 3, as shown in Figure 1(d).

We then partitioned metro stations into 5 and 3 groups for weekdays and weekends, respectively. Figure 1(b) shows the weekday activity patterns across time for each cluster. Cluster 1 reflects a stable bike usage across time. Cluster 2, 3, and 4 all represent a tidal activity patterns: for cluster 2, usage increases during afternoon and late evening peak hours; for cluster 3, most usage happens during morning peak hours; for cluster 4, activity spikes during afternoon peak. Cluster 5 exhibits a very distinct activity pattern that there is almost no activity during the day and most usage happens around midnight; we suspected that it is near a Jump hub and the activity is rebalancing around midnight. Similarly, Figure 1(e) shows the weekend activity patterns across time for each cluster. Compared to weekdays, the weekend activity has lower volumne in usage. Both cluster 1 and cluster 2 show a tidal activity pattern centering during morning and noon, respectively. Both cluster 3 and cluster 4 seem to occur due to rebalancing around early morning.

To better understand the relationship between spatial distribution of metro stations and the activity patterns of nearby dockless bikes, we plotted all metro stations, color-coded by clusters, on a DC city center map, as shown in Figure 1 (c) and 1 (f). Comparing these two sub-graphs, we realized that (1) not every metro station with nearby bike activity on weekdays also has activity on weekends; generally, people use dockless bike-sharing system more oftern on weekdays than on weekends. (2) There might be multiple Jump hubs for rebalancing the bikes and different hubs are used on weekdays and on weekends. (3) Metro stations associated with a tidal bike activity pattern are located either in city center (where people travel back from work in the evenings) or in the residential area (where people travel to work in the mornings).

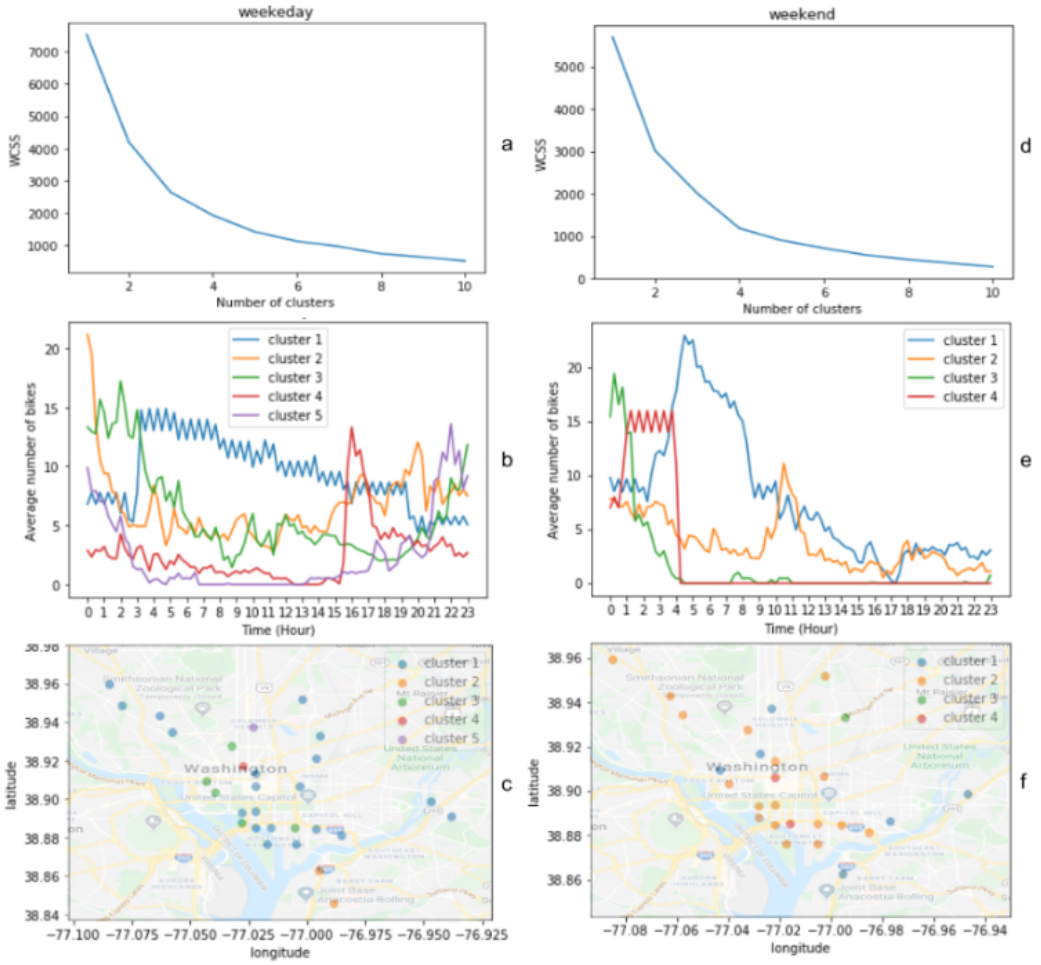

Figure 1: Cluster analysis for dockless bike-sharing system in DC. (a) Variation trend of partitioning error with k value on weekdays; (b) activity patterns of each cluster for metro stations on weekdays; (c) geographical distribution of each cluster for metro stations on weekdays; (d) variation trend of partitioning error with k value on weekends; (e) activity patterns of each cluster for metro stations on weekends; (f) geographical distribution of each cluster for metro stations on weekends.

3 Ridership Pattern of Stationed Bikes near Metro Stations: Case Study in DC, Boston and SF

3.1 Background

The high level goal of this project is to understand how each method within a large transportation system interacts. Particularly, we are interested in how the newly-emerging bike-sharing system interacts with the exciting and largely-used metro system. We hope to understand the role of bikes in the whole trip chain. Do people use bikes to commute from home/ workplace to the nearest metro station? Or do they completely replace metro with bikes when travelling for shorter distances? We hope to answer those questions using a data-driven approach with data for DC, Boston and SF transportation systems.

3.2 Data collection

3.2.1 Metro entry-exit data

Different from the previous section, other than name and location (latitude and longitude), this time we need more information about metro stations, including detailed entry-exit record across time. We will later compare the metro ridership pattern with the bike ridership pattern. We obtained relevant datasets from online sources published by the city governments. The data usually includes:

- Date
- Time period
- Entrance station
- Exit station
- Count

Particularly, we found entry-exit record during 15-minute time interval so that it is granular enough to conduct the comparison later.

3.2.2 Bike ridership data

The bike-sharing companies, Capital Bikeshare (DC), Bluebikes (Boston), and Bay Wheels (SF) provide detailed system data and publish downloadable files of trip history data each quarter. The standardized data file usually includes:

- Trip Duration (seconds)
- Start Time and Date
- End Time and Date
- Start Station ID
- Start Station Name
- Start Station Latitude
- Start Station Longitude
- End Station ID
- End Station Name
- End Station Latitude
- End Station Longitude
- Bike ID
- User Type (Subscriber or Customer – "Subscriber" = Member or "Customer" = Casual)

The data has been processed to remove trips that are taken by staff as they service and inspect the system, trips that are taken to/from any of our "test" stations at our warehouses and any trips lasting less than 60 seconds (potentially false starts or users trying to re-dock a bike to ensure it's secure).

3.3 Data pre-processing

3.3.1 Map metro station name to location

Since the entry-exit record obtained only includes the metro station name, we found additional station information with location and joined two datasets together. In this way, we mapped each metro station to its latitude and longitude, which will be used later to identify metro-bike pairs.

3.3.2 Identify metro-bike pairs

Similar to the previous section, we first filtered only data entries such that the origin or destination of a bike is a metro station. We considered bikes within 300 feet of a metro station as "near metro stations". After using latitudes and longitudes to calculate the distance between each metro station and bike station, we identify metro-bike pairs based on the above threshold.

3.3.3 Filter on weekdays only

We also kept only weekdays data since we expected very different ridership patterns on weekdays and weekends. Since we are interested in how people use metros and bikes for their daily commute, we decided to focus on weekday data.

3.3.4 Calculate average number of trips

We decided to focus on 15-minute time interval. For each time period, we computed the average number of trips starting and ending from a certain metro/ bike station.

3.4 Data analysis

3.4.1 Ridership analysis in DC

Metro Center and Chinatown are two metro stations with huge volumes. Therefore, we decided to focus on those two stations for the case study in DC. During the data pre-processing stage, we have already identified the nearby bike stations and computed the average number of trips starting and ending from those stations. To understand the similarity and difference between metro and bike ridership, we plotted the average number of trips for metros

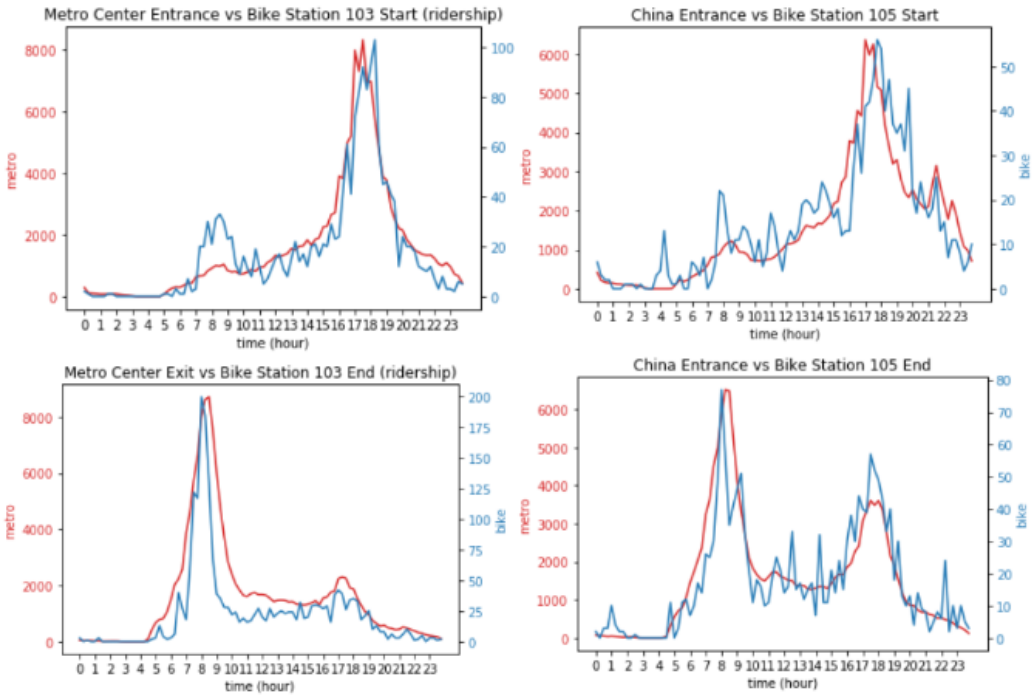

Figure 2: Ridership analysis for metro and bike-sharing systems in DC

and bikes (color-coded in red and blue, respectively) starting and ending from Metro Center and Chinatown, as shown in Figure 2.

We observed a very similar pattern in metro and bike ridership. For Metro Center as the starting point, people tend to use both metros and bikes a lot more during evening peak hours; for Metro Center as the destination, people use both metros and bikes a lot more during morning peak hours. Both metro and bike usages show a clear tidal pattern. Similarly, there is also a tidal pattern for Chinatown as the origin and destination; but it is more obvious that there are two peaks, one during morning rush hours and the other during the evenings.

3.4.2 Ridership analysis in Boston

Having observed a similar ridership pattern for metros and bikes in DC, we decided to look into other cities to see if the similarity is consistent across multiple cities. Therefore, we collected metro and bike data for Boston. As

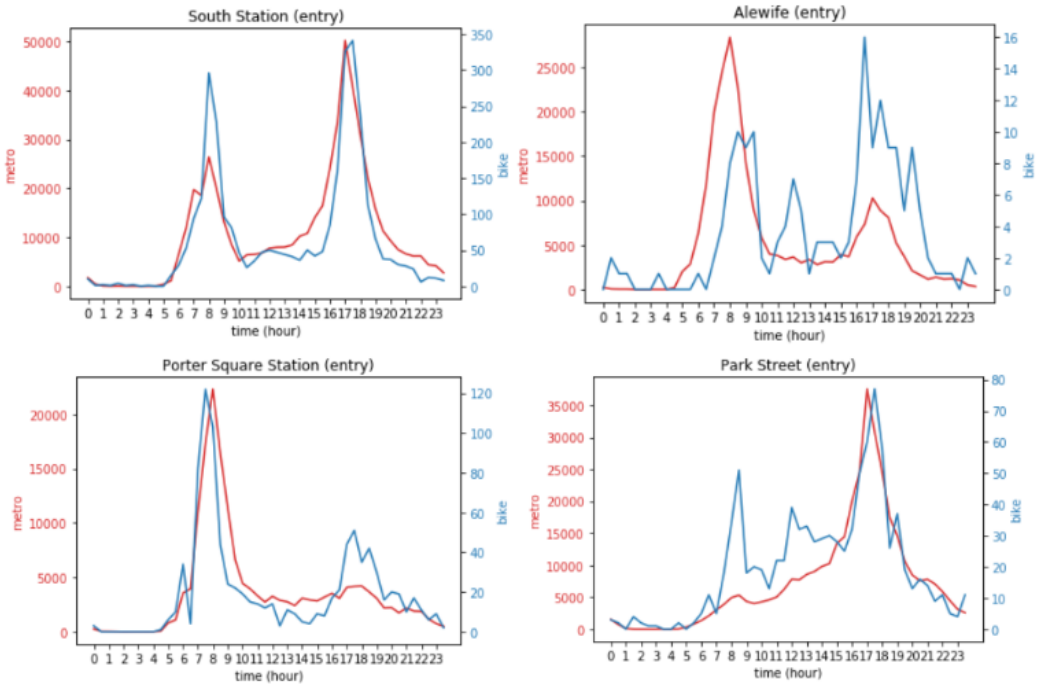

Figure 3: Ridership analysis for metro and bike-sharing systems in Boston

shown in Figure 3, South Station, Porter Square Station, and Park Street reflect a similar ridership pattern for metros and bikes, although the tidal patterns are different. For Alewife, however, the bike ridership pattern seems to be the opposite of the metro ridership pattern. We suspected that people near Alewife metro stations use metros to commute to work during the morning peak hours and use bikes to ride from metro station to home during the evening peak hours. This observation motivated us to conduct a even larger scale system-wide analysis in SF.

3.4.3 Ridership analysis in SF

The observation in ridership analysis for Boston raised another interesting question: will there be other relationship between the metro ridership and bike ridership patterns, other than the two (same and opposite) we have seen before? We decided to conduct a system-while analysis in SF by comparing the ridership patterns between BRAT and Bay Wheels.

There are 48 BART stations in SF and 18 of them have at least one nearby

bike station. Figure 4 shows the comparison between metro ridership and bike ridership for all 18 pairs as the starting and ending point, respectively. There are 36 sub-plots in total. It is obvious that there are more than two clusters of the pairs. We will further analyze the clusters in the next section.

3.5 Summary

Through case studies in DC, Boston, and SF, we concluded that there are various relationship between the metro ridership pattern and the bike ridership pattern. If two ridership patterns have the same trend, then it implies that people are more likely to use bikes as an alternative transportation method; people may choose to ride a bike instead of metro to go to places with shorter travel distance. If two ridership patterns have the opposite trend, however, then it implies that people are more likely to use bikes as a complementary method to the metro; people may use bikes to go from home or workplace to the nearby metro stations.

Ridership analysis in DC and Boston reveals stations with similar ridership patterns. Further analysis in Boston and SF shows that some stations may have opposite ridership patterns. A system-while analysis in SF shows that there may be more than two possible relationships between metros and bikes.

4 Cluster Analysis on Ridership of Stationed Bike-Sharing Systems near BART Stations: Case Study in SF

4.1 Background

In the previous section, we conducted a system-wide analysis for bike-sharing system near BART stations. It is obvious that we can partition stations into various groups based on the relationship between metro ridership and bike ridership.

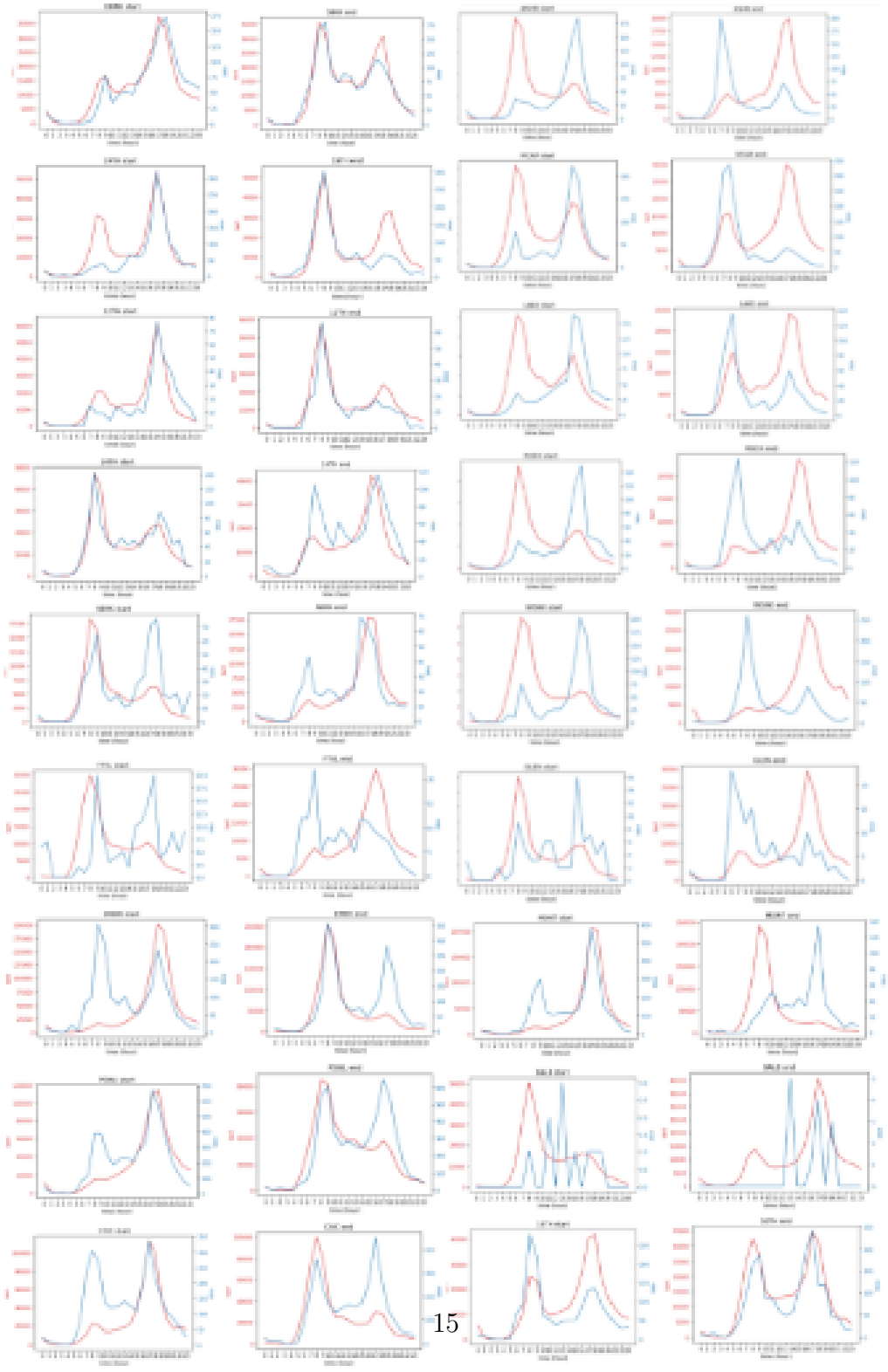

Figure 4: Ridership analysis for metro and bike-sharing systems in SF

4.2 Cluster analysis

As shown in Figure 5, cluster 1 (green) contains metro-bike pairs with same trend. Cluster 2 (red) contains metro-bike pairs with opposite trend. Cluster 3 (blue) contains metro-bike pairs such that there is a tidal pattern for metro ridership but there are two peaks in bike ridership; it implies that people use bikes during both morning and evening peak hours and there is no direction in the travel. Cluster 4 (yellow) contains three metro-bike pairs with some other trends different from the previous three.

To better understand how spatial distribution of stations influences the relationship between metro and bike ridership patterns, we plotted all 18 stations on the BART system map, as shown in Figure 6. We concluded that (1) BART stations with same ridership pattern (green) as bikes are usually in the business center, such as Downtown Berkeley, Oakland City Center, and 24th St Mission. In those areas, people are more likely to use bikes for short-distance commute. (2) BART stations with opposite ridership pattern (red) as bikes are usually in the residential area, such as Ashby, Rockridge, and Lake Merritt. People around those areas usually use bikes to travel between home or workplace to the nearby BART station. (3) BART stations in cluster 3 (blue) are around the downtown SF area, where people, including many tourists, use bikes for short-distance trips throughout the whole day. This explains the two peaks in bike ridership pattern.

5 Conclusion

Using a data-driven approach, we examined data about transportation systems in DC, Boston, and SF. In particular, we analyzed activity pattern of dockless bikes near metro stations and ridership pattern of stationed bikes near metro stations. Through various analysis approaches, including k-means clustering, we obtained a better understanding of the relationship between metro and bike usage. We concluded that there is a strong indication of the relationship between metro and bike ridership patterns by the spatial distribution of metro stations.

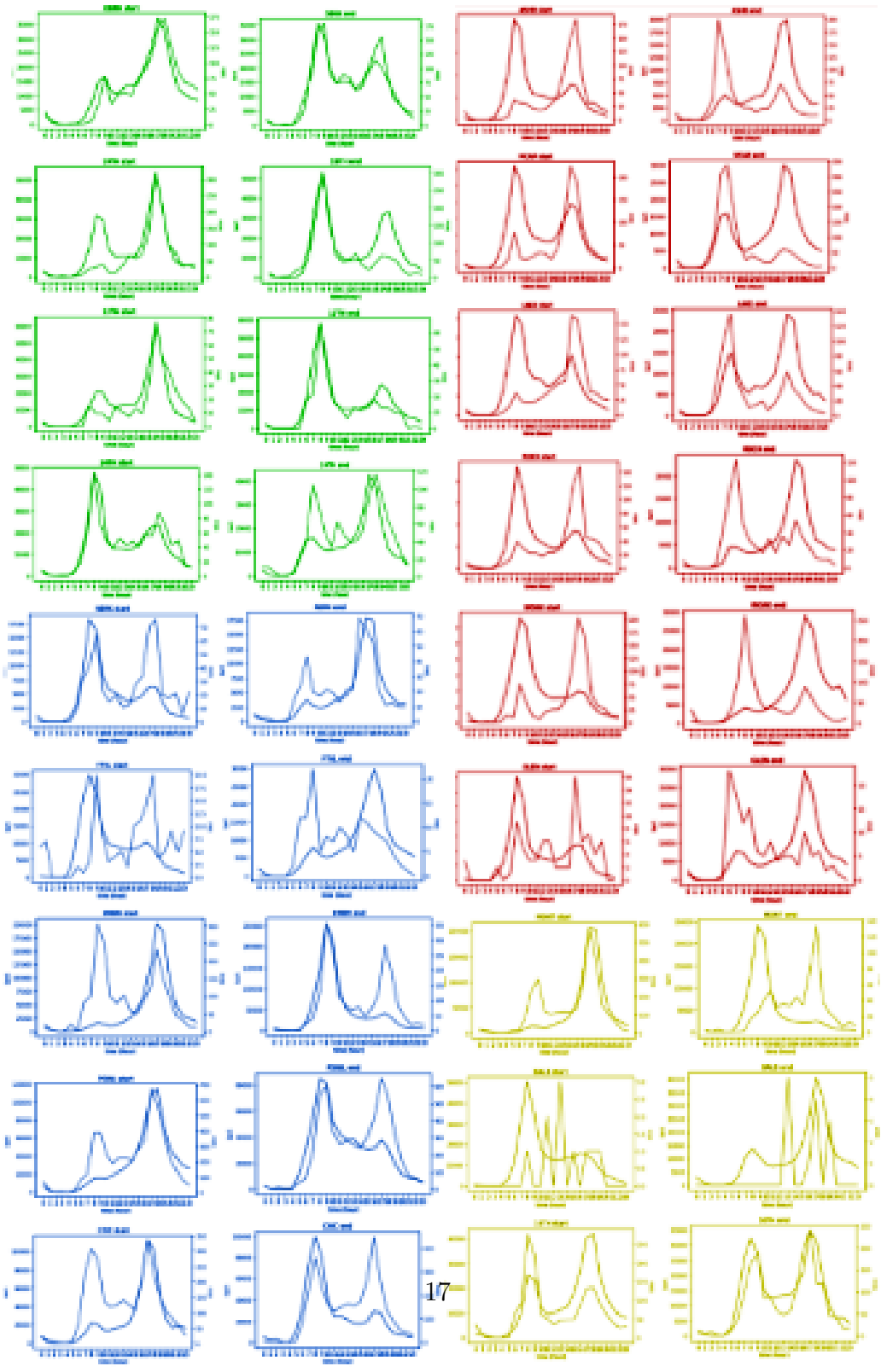

Figure 5: Ridership (clustered) for metro and bike-sharing systems in SF

Figure 6: Cluster analysis for metro and bike-sharing systems in SF