September 11th Memorial Program: Analysis of Spatial and Time of Day Subway Ridership Patterns in New York City

Jason Chen
Ph.D. Candidate
Civil Engineering
City College of New York
W. 140th Street and Convent Avenue
New York, NY 10031
Tel: 212-650-8290
Email: chenxq@ce.ccny.cuny.edu

Advisors:

Cynthia Chen, Ph.D.
Assistant Professor
Civil Engineering
City College of New York
W. 140th Street and Convent Avenue
New York, NY 10031
Tel: 212-650-5372
Email: c.chen@ccny.cuny.edu

and

James Barry
Manager, Transit Network Analysis
Division of Operations Planning
MTA New York City Transit
2 Broadway, Room A17.14
New York, NY 10004
Tel: 646-252-5631
Email: james.barry@nyct.com

Table of Contents

1. Project Overview

1-1. Introduction

Subway is the prevailing travel mode in New York City. In 2000, forty percent of the residents in New York City used subway as their daily commute mode. From 1990 to 2004, New York City's subway ridership has increased 48% (NYCT, 2004). In the mean time, rail transportation is facing a series of opportunities and challenges. According to the New York City population projection (Bloomberg and Burden, 2006), the city's population is expected to increase 13.9 percent from 2000 to 2030. This population increase is likely to put great pressure on the already crowded subways during peak time. Within the current capacity of the subway system, the additional ridership may be best accommodated during off-peak periods.

The existing literature on subway ridership focuses largely on the determinants of daily ridership (e.g. Al-Sahili and Taylor 1996; Bermello et al. 1997; Noland 2000; Syed and Khan 2001; Moore 2002). In these studies, the subject of interest is daily total ridership. The focus of the current study is time of day subway ridership pattern. In particular, we will examine how subway ridership evolves over a 24-hour day, develop a methodology to classify its time of day ridership pattern and a method to forecast the time of day ridership pattern for subway stations in New York City.

During the course of this project, we also investigated the spatial distribution of subway station ridership in New York City on weekdays and weekends. Since this part of the work is not closely related to our main focus: time of day ridership pattern in New York City, we report this part of the analysis in Appendix.

1-2. Research Objectives

In this study, we expect to answer three sets of questions:

1. Time of day ridership pattern
    a. How does ridership distribute over 24 hours for different stations?
    b. What specific patterns can we identify?
2. Time of day ridership pattern and local features
    a. How is the time of day ridership pattern related to various factors, including local area effects (e.g., population and employment in the local area, and land use), and network position effects (e.g., general travel cost to CBD)?
    b. What conclusions can we draw?
3. Predicting time of day ridership pattern
    a. How can we reliably predict time of day ridership pattern for a station?

1-3 Project Datasets

The research is made possible with the financial support of September 11th Memorial Program. Professor Cynthia Chen of City College of New York serves as a faculty advisor and Mr. James Barry of New York City Transit serves as a professional advisor.

The main dataset used in the study is the average subway ridership data every 6 minutes for one day[1] in May, 2005. In addition, we obtained datasets on subway routes, subway stations, socio-economic and demographic data, employment data, and land use dataset. In detail, these datasets as well as their sources are listed as follows:

- Average weekday, Saturday, and Sunday ridership data for May 2005 by 6 minute interval (from New York City Transit),
- Average weekday, Saturday, and Sunday transfer ridership (bus to subway) information for May 2005 by one hour interval (from New York City Transit),
- Subway route, bus route, subway station and bus station files in ArcGIS format (from New York City Transit),
- New York City Street map files (from ArcGis America Street Map File),
- Land use data (from PLUTO file released in 2006),
- Socio-economic and demographic data in Census Tract level (from US 2000 census),
- Employment information by census tract (from Census Transportation Planning Package),
- Station network position information[2] (measured as the general travel cost to there zones[3] in CBD, from New York City Transit).

Figure 1-1 shows all the subway stations in New York City. In all, there are 542 stations in New York City. Among them, 165 stations are in Manhattan; 186 stations are in Brooklyn; 95 stations are in Queens; 74 stations are in Bronx; and 22 stations are in Staten Island. Stations in Staten Island are excluded from the study, due to unavailability of the ridership data on these stations[4].

The report is organized as follows. In Chapter 2, we describe various time of day ridership patterns. Its connection with local land use, socio-economic and demographic characteristics, and the relative position of a station in the subway network are discussed in Chapter 3. In Chapter 4, we describe our methodology to use daily total ridership and ridership pattern information (described in Chapter 3) to forecast a station’s time of day ridership pattern. The conclusions follow in Chapter 5. At the end of Chapter 5, we provide two appendices. In Appendix, we present results on the spatial distribution of the ridership data.

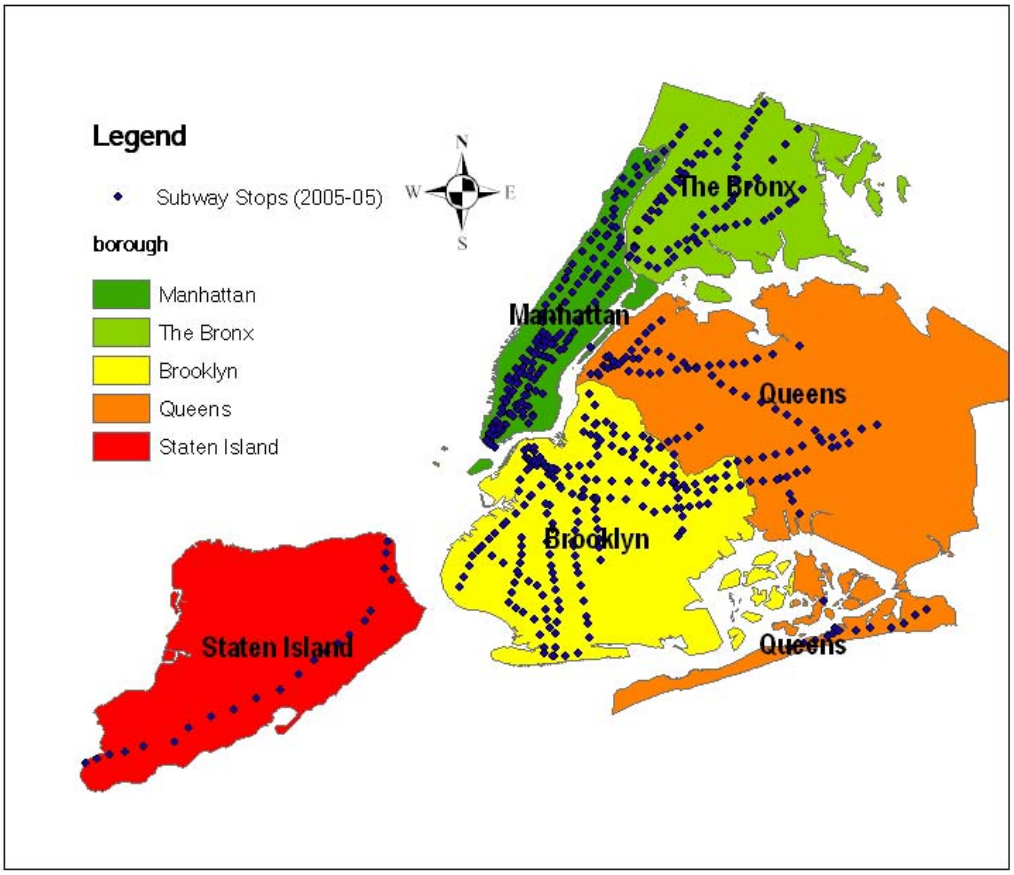
[1] This one day is the average day where ridership in each 6-minute interval throughout the month of May is averaged.

[2] This is measured by the general travel cost to CBD.

[3] CBD is divided into three parts: midtown, downtown, and valley.

[4] Stations in Staten Island are not run by New York City Transit.

Figure 1-1: Subway Stations in New York City



2. Time of Day Ridership Pattern – When Do They Occur?

2-1. Observations from Stations along Line 7

The objective of our preliminary analysis of time of day ridership pattern is to observe time of day patterns of stations along subway line number 7. Stations along line 7 include stations locating in CBD (e.g. 42nd street Times Square), stations locating in areas with high transfer ridership and many Asian immigrants (e.g. Flushing, Main St.), stations locating in recreational areas (Shea Stadium), and stations locating in residential areas (e.g. Bliss St.). It is thought that the variety of land use around number 7 stations will provide us insights in time of day ridership pattern.

Line 7 starts from Times Square station, which is the core of Manhattan CBD, and ends at Flushing Main St., where a substantial share of the residents are Asian immigrants. Main street station in Flushing is located in the joint area of two zipcodes: 10034 and 10035. In the area with zipcode 10034, 43% of the population is Asian; the next group is white alone, which is 41%. In the area with zipcode 10035, 54% of the population is Asian. The next group, White alone, takes about 27%. Along the line, there are nineteen stations. These stations locate in various types of land use areas – residential, commercial, or mixed land use. The average daily riderships of these stations for weekday and weekend are presented in Table 2-1.

Table 2-1 Average Daily Ridership for Stations along Line 7

Station Name	**Weekday**	**Saturday**	**Sunday**	**Ratio_1***	**Ratio_2***	**Ratio_3***
Main St.	56,770	36,152	27,728	64%	49%	77%
Shea Stadium	5,051	7,442	7,702	147%	152%	103%
111St.	9,391	7,636	6,336	81%	67%	83%
Corona Plaza	15,156	12,357	10,031	82%	66%	81%
Junction BLVD	20,439	16,498	12,851	81%	63%	78%
Elmhurst Av.	16,912	13,793	11,448	82%	68%	83%
Jackson Hts.	17,426	14,223	12,198	82%	70%	86%
69 St.	5,087	4,092	3,659	80%	72%	89%
61 St.	15,645	10,713	8,709	68%	56%	81%
52 St.	6,634	4,689	3,797	71%	57%	81%
Bliss St.	14,075	10,431	8,149	74%	58%	78%
Lowery	10,484	6,549	5,008	62%	48%	76%
Rawson	13,804	2,996	1,066	22%	8%	36%
Court House Sq.	9,059	5,561	3,654	61%	40%	66%
Hunters point Av.	6,468	1,141	813	18%	13%	71%
Jackson Av.	6,597	3,018	2,244	46%	34%	74%
Queensboro Plaza	8,855	4,833	3,198	55%	36%	66%

Station Name	Weekday	Saturday	Sunday	Ratio_1*	Ratio_2*	Ratio_3*
74 St/Broadway	44,135	34,174	28,016	77%	63%	82%
Times Square	173,260	118,257	96,352	68%	56%	81%

* Ratio_1= Saturday daily ridership /weekday daily ridership
Ratio_2= Sunday daily ridership /weekday daily ridership
Ratio_3= Sunday daily ridership/Saturday daily ridership

For stations along line 7, weekday riderships are always higher than that on weekends, except Shea Stadium. For fourteen out of nineteen stations, their Saturday daily riderships decrease to about 60% to 82% of weekday ridership; Sunday riderships reduce to about 40% to 72% of weekday ridership (see Table 2-1). At the Shea Stadium station, the highest daily ridership occurs on Sunday, which is probably due to recreational activities at the Stadium. Riderships at Rawson station and Hunters point Av. station experience a significant reduction on Saturday (80% reduction) and Sunday (90% reduction). This is probably due to that the land use around these two stations is primarily office buildings.

In the following section, we discuss time of day patterns for weekday, Saturday and Sunday separately.

Observed time of day ridership patterns for weekday

Peak starting time and peak volume are two important attributes. Table 2-2 presents morning and afternoon peak starting and ending time, and weekday volume for each station. In general, morning peak starting and ending time fall in the period of 7:00~9:00 a.m. and afternoon peak starting and ending time fall in the period of 3:00~5:00 p.m. In sum, we make the following observations:

a. Fourteen of the nineteen stations along subway line number 7 have a morning peak volume that is higher than afternoon peak volume. For these stations, the ratio of morning peak volume over afternoon peak volume ranges from 1.53 to 4.13. The morning peak starting time of these stations ranges from 7:00 to 8:00 a.m. The afternoon peak starting time of these stations ranges from 3:00 to 5:00 p.m. Assume that work trips account for the majority of morning peak volume and home trips account for most of the afternoon peak volume, the ratio of morning peak volume over afternoon peak volume reflects the land use patterns of the areas around stations. According to the ratios, we further separate these fourteen stations into two group:
    i. Six stations (111 St., Corona Plaza, Junction BLDV, Elmhurst AV., 61 St., 52 St.) have ratios over 3.0 (see Table 2-2), which may indicate that they locate in mostly residential areas.
    ii. Eight stations (Jackson Hts., Lowery, Hunters point AV., Main St., 69 St., Bliss St., Jackson AV., 74 St/Broadway) have ratios between 1.53 and 3.0, which may indicate that they locate in mixed use areas.
b. Three stations (Court House Sq., Queensboro Plaza, Times Square) have ratios of morning peak volume over afternoon peak volume under 1.0, which may indicate that they locate in commercial areas. The afternoon peak starting time for Court House Sq. station is from 5:00.p.m., and its weekend afternoon peak volume is only 37% (473, see Table 2-3) and 23% (292, see Table 2-4) of weekday afternoon peak ridership (1,278, see Table 2-2) for

Saturday and Sunday respectively. This suggests that Court House Sq. station locates in area with a high share of commercial land use. This likely applies to Queensboro Plaza as well. Times Square is different. Times Square station has a night peak on weekends, suggesting a high share of recreational land use around the Times Square station.

c. For Shea Stadium, the highest peak volume occurs in the evening, which may be the result of recreational activities in the stadium conducted at night.
d. Rawson station doesn't have a morning peak but have an afternoon peak. This may suggest more of a commercial land use than a residential one around the station. It could also be possible that many residents there are working at night shift.
e. Times Square and Main St. stations have a high morning peak volume. Due to the fact that they locate in highly commercial or mixed use areas, the high ridership may not come from local residents but from people who transfer there. Times Square has no morning peak during weekends. The Main St. station has a very long peak time during weekends (see Table 2-3 and Table 2-4).

Table 2-2 Weekday Peak time Volume Summary

Station Name	Morning Peak (a.m.)	Morning Peak Volume	Afternoon Peak (p.m.)	Afternoon Peak volume	Evening Peak (p.m.)	Evening Peak Volume	Ratio*
Main St	8:00-9:00	8,741	5:00-6:00	3,453	no	no	2.53
Shea Stadium	8:00-9:00	285	4:00-5:00	455	9:00-10:00	1051	0.63
111St	7:00-8:00	1,609	4:00-5:00	390	no	no	4.13
Corona Plaza	7:00-8:00	2,382	3:00-4:00	662	no	no	3.60
Junction BLVD	7:00-8:00	3,529	3:00-4:00	900	no	no	3.92
Elmhurst Av	7:00-8:00	2,724	5:00-6:00	774	no	no	3.52
Jackson Hts	7:00-8:00	1,792	5:00-6:00	1,174	no	no	1.53
69 St	7:00-8:00	687	5:00-6:00	293	no	no	2.34
61 St	8:00-9:00	2,411	5:00-6:00	783	no	no	3.08
52 St	8:00-9:00	1,105	3:00-4:00	302	no	no	3.66
Bliss St	8:00-9:00	2,065	5:00-6:00	778	no	no	2.65
Lowery	8:00-9:00	1,534	4:00-5:00	812	no	no	1.89
Rawson	no	no	5:00-6:00	2,015	9:00-10:00	801	no
Court House Sq	8:00-9:00	732	5:00-6:00	1,278	no	no	0.57
Hunters point Av	8:00-9:00	1,415	4:00-5:00	780	no	no	1.81
Jackson Av	8:00-9:00	1,059	5:00-6:00	504	no	no	2.10
Queensboro Plaza	8:00-9:00	434	5:00-6:00	1,190	no	no	0.36
74 St/Broadway	8:00-9:00	5,957	4:00-5:00	2,486	no	no	2.40
Time Square	8:00-9:00	13,687	5:00-6:00	19,244	no	no	0.71

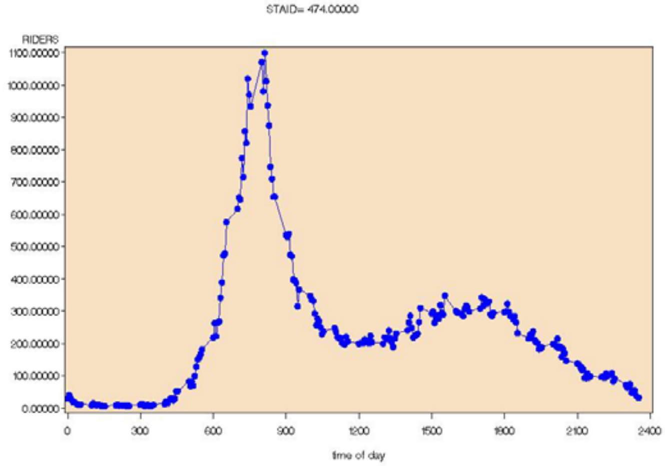
*Ratio=Morning Peak Volume/Afternoon Peak Volume

Visually, we can divide weekday time of day ridership patterns into four distinct categories: high morning peak pattern, high afternoon peak pattern, no morning peak pattern, and evening peak pattern. The features of each pattern as well as its associated graphical plot are presented below.

(1) Pattern 1: high morning peak pattern

In this pattern, morning peak volume is higher than afternoon peak volume. A graphic plot for this pattern is presented in Figure 2-1. Stations that belong this pattern include: main St., 111 St., Corona Plaza, Junction BLDV., Elmhurst Av., Jackson Hts., 69 St., 61 St., 52 St., Bliss St., Lowery, Hunter Point Av., Jackson AV., and 74 St./Broadway.

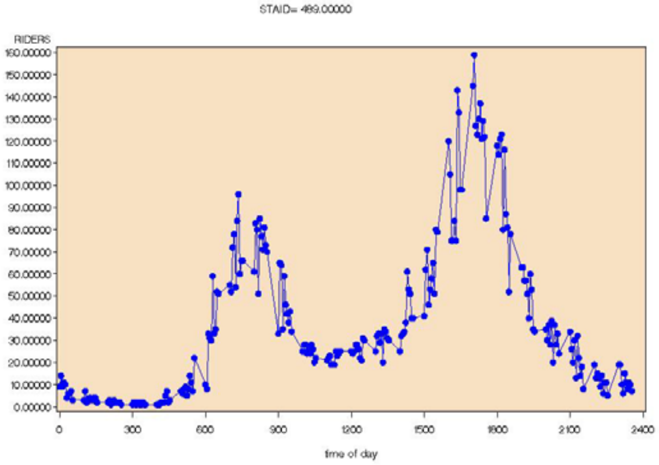
Figure 2-1 Sample Temporal Pattern 1 in Weekday (Main St.)



(2) Pattern 2: high afternoon peak pattern

In pattern 2, afternoon peak volume is higher than morning peak. A graphical plot for pattern 2 is presented in Figure 2-2. Stations with this pattern include: Court House Sq., Queensboro Plaza., and Times Square.

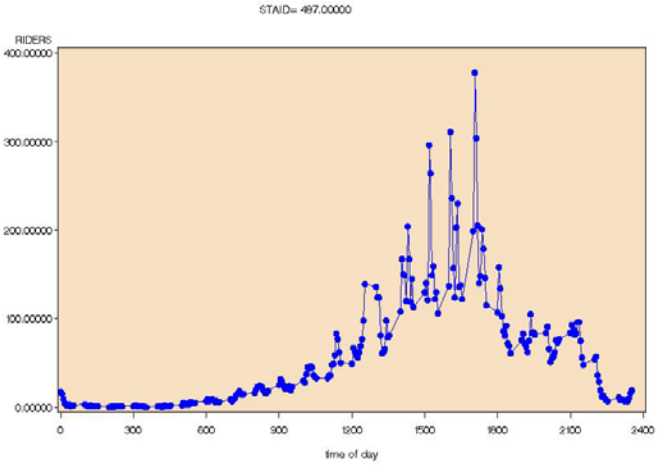
Figure 2-2 Sample Temporal Pattern 2 in Weekday(Court House Sq.)



(3) Pattern 3: no morning peak pattern

The main feature of pattern 3 is its lack of a morning peak. Most of the riders enter the stations in the afternoon. A graphical plot for pattern 3 is presented in Figure 2-3. Rawson station has such a pattern. Recall that the Rawson station has a significant reduction in daily ridership from weekday to weekend. Both time of day pattern and the change in daily ridership between weekday and weekend indicate that Rawson station probably locates in areas with heavily commercial land use.

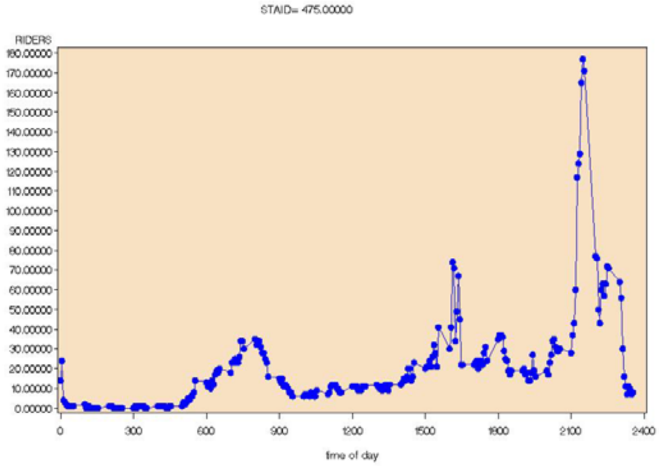
Figure 2-3 Sample Temporal Pattern 3 (Rawson)



(4) Pattern 4: evening peak pattern

The existence of an evening peak is probably due to some recreational activities around the station. Only Shea Stadium station has such a pattern, which is presented in Figure 2-4. Recall that the ridership in Shea Stadium station increases on weekends. Both the evening peak and increased ridership on weekend imply that the night recreational activities associated with Shea Stadium station.

Figure 2-4 Sample Temporal Pattern 4 (Shea Stadium)



Observed time of day ridership patterns for Saturday:

Morning and afternoon peak starting and ending time and volume for Saturdays are presented in Table 2-3. Compared to weekday, the Saturday pattern has several unique features. In some stations, morning peak disappeared. The ratio of morning peak volume over afternoon peak volume becomes smaller than that on weekday. The starting time of afternoon peak also becomes earlier than that of weekday. From Table 2-3, we can make several observations:

a. Seven stations -111 St., Junction BLDV, Elmhurst AV, 69 St., 61 St., Hunters point AV., Queensboro Plaza, 74 St./Broadway – have morning peak and afternoon peak on Saturday. The ratio of morning peak volume over afternoon peak ranges from 1.91 ~ 0.93, which is much smaller than that of weekday and is likely resulted from the reduced work trips on Saturday.
b. Stations “Main St., Bliss St., and Lowery” have very long peak duration (Main St: from 8:00 am.to 6:00 pm.; Bliss St: from 8:00 am. to 4:00 pm.; Lowery: from 8:00 am. to 4:00 pm.).
c. Stations “Shea Stadium, Jackson Hts., Rawson, Court House Sq” have afternoon peaks only, which indicates a mostly commercial land use pattern around stations.
d. Times Square station has both an afternoon and an evening peak, which suggests a mostly commercial and recreational land use pattern.

Table 2-3 Saturday Peak Time Volume Summary

Station Name	**Morning Peak (am.)**	**Peak Volume**	**Afternoon Peak (pm.)**	**Peak volume**	**Night Peak (pm.)**	**Peak Volume**	**Ratio**
Main St	8:00 am.-6:00 pm.	2,400	No	No	No	No	
Shea Stadium	No	No	4:00-5:00	4,121	No	No	
111St	7:00-8:00	781	4:00-5:00	409	No	No	1.91
Corona Plaza	7:00-8:00	1,255	No	No	No	No	
Junction BLVD	7:00-8:00	1,672	2:00-3:00	993	No	No	1.68
Elmhurst Av	7:00-8:00	1,283	3:00-4:00	808	No	No	1.59
Jackson Hts	No	No	2:00-3:00	1,112	No	No	
69 St	9:00-10:00	296	2:00-3:00	274	No	No	1.08
61 St	8:00-9:00	839	12:00-1:00	820	No	No	1.02
52 St	9:00-10:00	384	No	No	No	No	
Bliss St	8:00-16:00	749	No	No	No	No	
Lowery	8:00-16:00	471	No	No	No	No	
Rawson	No	No	2:00-3:00	381	No	No	
Court House Sq	No	No	5:00-6:00	473	No	No	
Hunters point Av	9:00-10:00	105	4:00-5:00	87	No	No	1.21
Jackson Av	No	No	12:00-1:00	281	No	No	
Queensboro Plaza	12:00-13:00	538	4:00-5:00	427	No	No	1.26

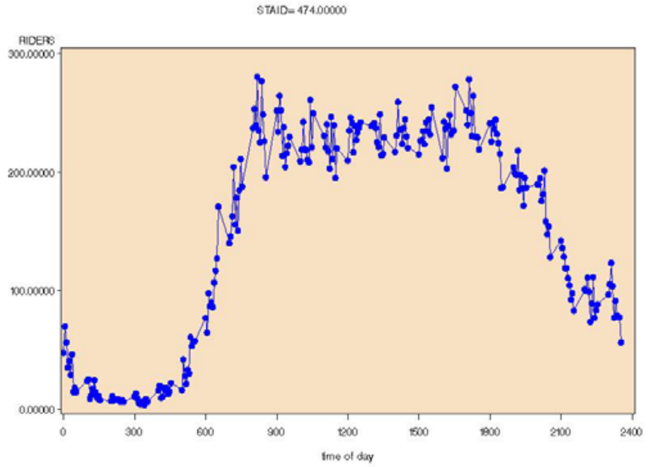
Station Name	**Morning Peak (am.)**	**Peak Volume**	**Afternoon Peak (pm.)**	**Peak volume**	**Night Peak (pm.)**	**Peak Volume**	**Ratio**
74 St/Broadway	9:00-10:00	2,306	3:00-4:00	2,474	No	No	0.93
Times Square	No	No	4:00-5:00	9,281	10:00-11:00	8384	

In sum, we can categorize Saturday time of day patterns into six groups. The features of each group and its associated graphical plot are presented below.

(1) Pattern 1: One long duration peak pattern

In this pattern, the peak starts in the morning and lasts until afternoon, which is shown in Figure 2-5. Stations that have this pattern include Main St., Jackson Hts., 69 St., Bliss St., Lowery., 74 St./Broadway.

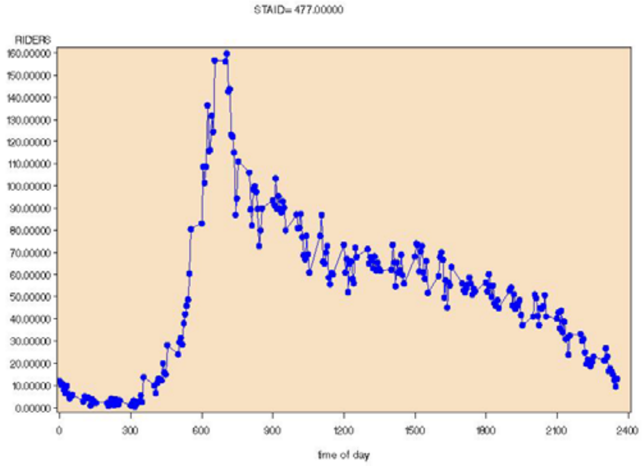
Figure 2-5 Sample Temporal Pattern 1 (Main St.)



(2) Pattern 2: Morning peak only pattern

There is only a morning peak in the 24-hour period in pattern 2, which is presented in Figure 2-6. Stations that have this pattern include Corona Plaza, Junction BLDV, 61 St., 52 St.

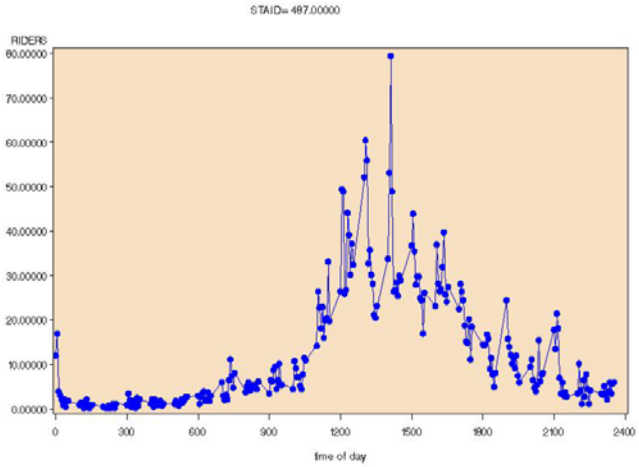
Figure 2-6 Sample Temporal Pattern 2 (Corona Plaza)



(3) Pattern 3: Afternoon peak only pattern

Stations with this pattern include Shea Stadium and Rawson stations. They only have an afternoon peak in the 24-hour period. The pattern is presented in Figure 2-7.

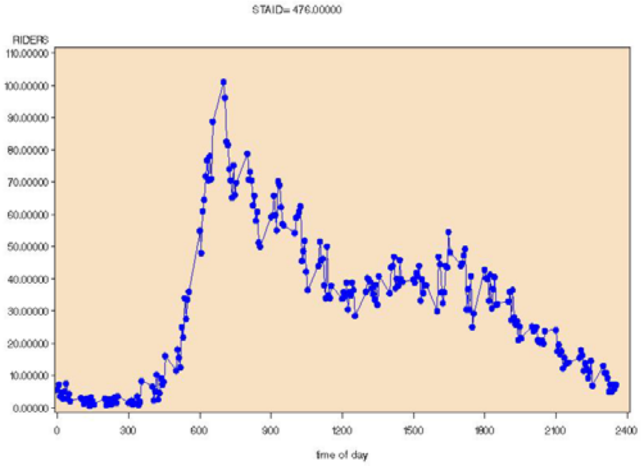
Figure 2-7 Sample Temporal Pattern 3 (Rawson)



(4) Pattern 4: High morning peak pattern

Stations with this pattern include 111 St., Elmhurst Av., and Hunter Point Av. stations. In this pattern, morning peak volume is higher than afternoon peak volume, which is present in Figure 2-8.

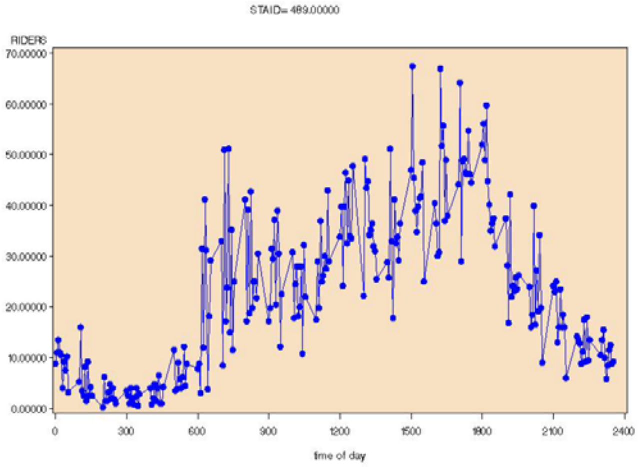
Figure 2-8 Sample Temporal Pattern 4 (111 St.)



(5) Pattern 5: High afternoon peak pattern

Only Court House Sq. station has this pattern. Shown in Figure 2-9, afternoon peak volume is higher than morning peak.

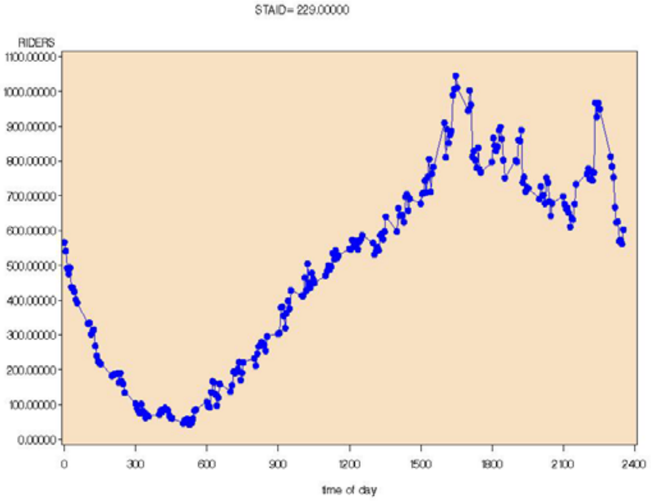
Figure 2-9 Sample Temporal Pattern 5 (Court House Sq.)



(6) Pattern 6: Evening peak pattern

Times Square station has this pattern. Presented in Figure 2-10, this pattern has no morning peak, but an evening peak.

Figure 2-10 Sample Temporal Pattern 6 (Times Square)



Observed patterns for Sunday:

Morning and afternoon peak starting and ending time and volume on Sunday are presented in Table 2-4. Compared to Saturday and weekday, Sunday's morning peak starting time is later and its afternoon peak time is earlier. Many stations don't have a morning peak. From Table 2-4, we could observe some distinguishing features for Sunday.

a. Sunday has fewer stations (Only three stations "111 St., Elmhurst AV., Hunters point AV") that have morning and afternoon peak, and the ratio of morning peak over afternoon peak volume is smaller than that of weekday and Saturday.
b. Seven stations (Main St., Shea Stadium, Junction BLDV, Jackson Hts, Bliss St., Rawson, Court House Sq., Queensboro Plaza, and 74 St/Broadway) don't have a morning peak but an afternoon peak.

Table 2-4 Sunday Peak Time Volume Summary

Station Name	**Morning Peak (am.)**	**Peak Volume**	**Afternoon Peak (pm.)**	**Peak Volume**	**Night Peak (pm.)**	**Peak Volume**	**Ratio**
Main St	No	No	1:00-4:00	1,996-2,069	No	No	
Shea Stadium	No	No	4:00-5:00	2,269	No	No	
111St	10:00-11:00	512	4:00-5:00	461	No	No	**1.11**
Corona Plaza	9:00 am.-4:00 pm.	690-736	No	No	No	No	
Junction BLVD	No	No	2:00-3:00	1,034	No	No	
Elmhurst Av	10:00-11:00	865	2:00-3:00	858	No	No	**1.01**
Jackson Hts	No	No	2:00-3:00	1,173	No	No	
69 St	12:00 pm.-1:00 pm.	335	2:00-3:00		No	No	

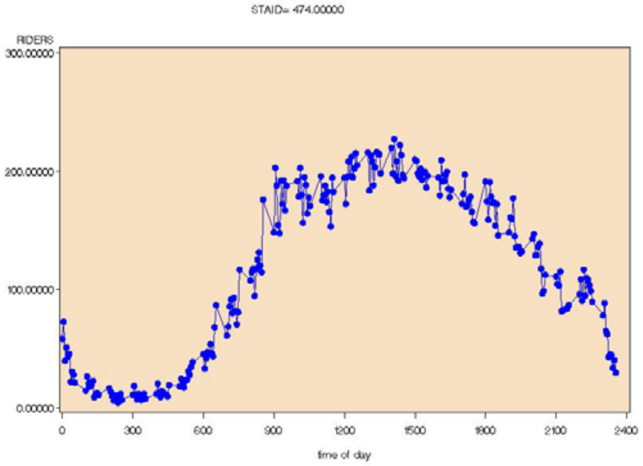
Station Name	**Morning Peak (am.)**	**Peak Volume**	**Afternoon Peak (pm.)**	**Peak Volume**	**Night Peak (pm.)**	**Peak Volume**	**Ratio**
61 St	11:00-12:00	729	12:00-13:00		No	No	
52 St	9:00 am.-3:00 pm.	296-311	No	No	No	No	
Bliss St	No	No	3:00-4:00	664	No	No	
Lowery	11:00 am.-2:00 pm.	389-409	No	No	No	No	
Rawson	No	No	4:00-5:00	142	No	No	
Court House Sq	No	No	6:00-7:00	292	No	No	
Hunters point Av	9:00-10:00 11:00-12:00	72 78	4:00-5:00	78	No	No	0.92
Jackson Av	12:00-14:00	216~218	No	No	No	No	
Queensboro Plaza	No	No	5:00-6:00	285	No	No	
74 St/Broadway	No	No	2:00-3:00	2,291	No	No	
Times Square	No	No	5:00-6:00	7,897	9:00-10:00	6585	

In sum, we can categorize the Sunday time of day patterns into four groups. The features of each group and its associated plot are discussed below.

(1) Pattern 1: One long duration peak pattern

Stations with this pattern include Main St., 111 St., Corona Plaza., Junction BLDV., Elmhurst Av., 69 St., 61 St., 52 St., Bliss St., Lowery, Court House Sq., and 74 St./Broadway. Shown in Figure 2-11, the outstanding feature of this pattern is that the "peak" lasts for several hours.

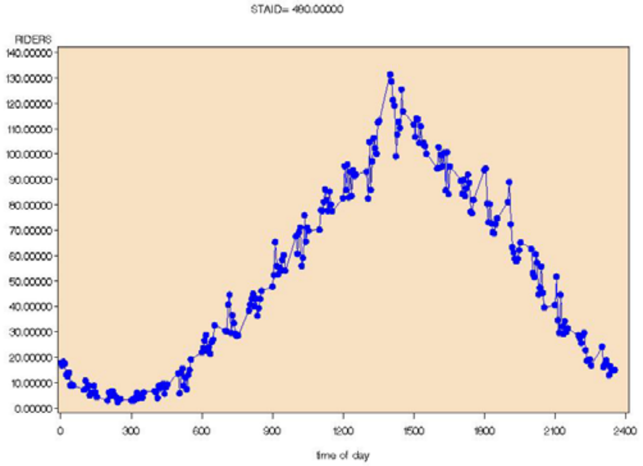
Figure 2-11 Sample Temporal Pattern 6 (Main St.)



(2) Pattern 2: Only one afternoon peak

Stations with this pattern include Shea Stadium, Jackson Hts., Rawson, Jackson AV., and Queensboro Plaza. There is only an afternoon peak during the 24-hour period in this pattern, which is presented in Figure 2-12..

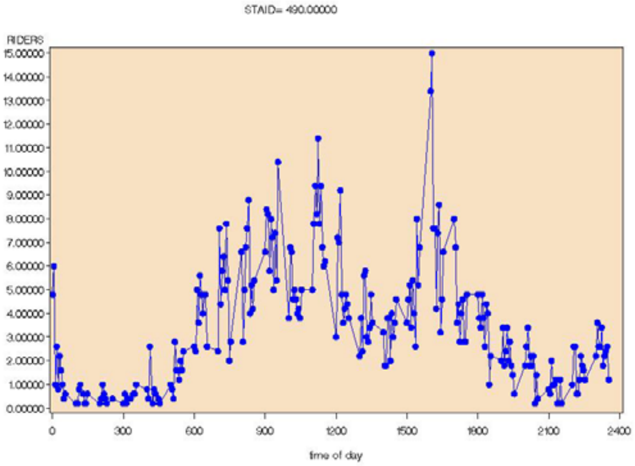
Figure 2-12 Sample Temporal Pattern (Jackson Hts.)



(3) Pattern 3: Morning peak and afternoon peak pattern

Shown in Figure 2-13, patten 3 has both a morning peak and an afternoon peak. Only Hunter Point Av. station has this pattern.

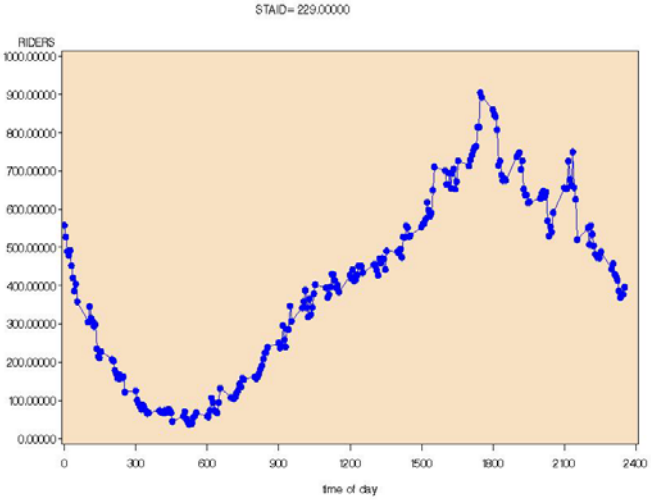
Figure 2-13 Sample Temporal Pattern (Hunter Point Av.)



(4) Pattern 4: Afternoon peak and evening peak

Stations in this pattern have both an afternoon peak and an evening peak. Only Times Square station has such pattern, which is shown in Figure 2-14.

Figure 2-14 Sample Temporal Pattern (Times Square)



Summary

We can make several conclusions. First, time of day pattern can be characterized by many attributes, including peak volume, peak starting and ending time, and ratio of morning peak volume over afternoon peak volume. Second, there is a significant difference between weekday, Saturday and Sunday pattern. Compared to weekday, ridership on weekends is more evenly distributed during the day. Third, the built environment is likely to be associated with time of day ridership pattern.

2-2. Ridership Time of Day Distribution for All Stations by Cluster Analysis

Variable list

Upon lengthy discussions with NYCT, we created nineteen variables (Table 2-5) to describe the various attributes of time of day ridership pattern. These nineteen variables can be categorized into four groups: concentration of riderhship, position of concentration in time, transfer ridership,, and relative magnitude of concentration in different time periods.

Table 2-5 Variables Used in Cluster Analysis

Variable Name	**Note**
Position of Concentration in time	
Starting time of morning peak	The starting time is measured at 6 minutes interval, e.g. 8:00, 8:06, 8:12 etc. The latest morning peak starting time is set as 10:00 a.m. If there is no peak until that time, no morning peak is recorded.
Width of morning peak	The width of peak is measured at 30 minutes interval, e,g. 60 minutes, 90 minutes etc.
Starting time of afternoon peak	The starting time is measured at 6 minutes interval.
Width of the afternoon peak	The width of peak is measured at 30 minutes interval, e.g. 60 minutes, 90 minutes etc.
Concentration of ridership	
Morning peak hourly ridership	A morning peak period is defined such that the average ridership (in half an hour) before and after this peak is less than 90% of peak hour, which is the hour with the maximum ridership in the peak period.
Afternoon peak hourly ridership	An afternoon peak is defined such that the average ridership (in half an hour) before and after this peak is less than 90% of peak hour, which is the hour with the maximum ridership in the peak period.
Early morning hourly ridership	The period between midnight and the starting time of morning peak is defined as early morning period. The hourly ridership is the average ridership per hour during this period.

Variable Name	Note
Midday hourly ridership	The period between the end of morning peak and the starting time of afternoon peak is defined as midday period. The hourly ridership is the average ridership per hour during this period.
Transfer ridership ratio (bus to subway)	
Ratio of transfer ridership in early morning ridership	This is measured by the ratio of hourly transfer ridership (from bus to subway) over hourly early morning ridership
Ratio of transfer ridership in morning peak ridership	This is measured by the ratio of hourly transfer ridership (from bus to subway) over hourly morning peak ridership
Ratio of transfer ridership in midday ridership	This is measured by the ratio of hourly transfer ridership (from bus to subway) over hourly midday ridership
Ratio of transfer ridership in afternoon peak ridership	This is measured by the ratio of hourly transfer ridership (from bus to subway) over hourly afternoon peak ridership
Relative magnitude of ridership	
Ratio of morning peak hourly ridership over total daily ridership	This is measured by the morning peak hourly ridership divided by total daily ridership
Ratio of afternoon peak hourly ridership over total daily ridership	This is measured by the afternoon peak hourly ridership divided by total daily ridership
Ratio of morning peak hourly ridership over afternoon peak hourly ridership	This is measured by the morning peak hourly ridership divided by afternoon peak hourly ridership
Ratio of early morning hourly ridership over total daily ridership	This is measured by the early morning hourly ridership divided by total daily ridership
Ratio of midday hourly ridership divided by total daily ridership	This is measured by the midday hourly ridership divided by total daily ridership
Ratio of early morning hourly ridership over maximum hourly ridership	This is measured by the early morning hourly ridership divided by maximum hourly ridership
Ratio of midday hourly ridership over maximum hourly ridership	This is measured by the midday hourly ridership divided by maximum hourly ridership

Cluster Procedure

Since each station is associated with a set of attributes described in Table 2-5, cluster analysis is used to classify stations into clusters each with a distinctive pattern. Whether two stations belong to the same cluster is determined by the Euclidean Distance between them. Let p_1 be the standardized variable 1 for station P, such as morning peak hourly ridership, and p_2 be the standardized variable 2 for station P, such as afternoon peak hourly ridership. Station P can be represented in a K-dimension space as: P = (p_1,p_2,…p_K) and the same can apply to station Q. The Euclidean Distance between stations P and Q is defined as: $\sqrt{\sum_{i=1}^{K}(p_i - q_i)^2}$.

The underlying principle of cluster analysis is to identify group membership for each station, such that the sum of the Euclidean Distances within a group is minimized and the sum between groups is maximized. To achieve this criterion, K-means algorithm is used. K-means cluster analysis starts with a user-identified number of clusters (it should be significantly large) and the same number of randomly picked centroids serving as cluster means. An individual observation is compared with the value of each centroid and assigned to the cluster to whom it has the shortest distance. Each centroid is updated after a new assignment. The process is completed after a complete pass through the dataset. Then, the process is repeated again, but with the centroid achieved in the previous run. In other words, after the first run, the centroids are no longer randomly picked. After several runs, the result would converge to an optimum solution, in which the distance between clusters is maximized and the distance within a cluster is minimized.

The cluster analysis with K-means algorithm is preceded by the FASTCLUS procedure of SAS software. By default, every station is a cluster seed or cluster center in FASTCLUS procedure. In FASTCLUS procedure, the maximum number of clusters is defined by researcher. Then the FASTCLUS procedure grouped those stations that are close to each other in Euclidean distances into one cluster until the total number of clusters reaches its maximum. In this way, stations with similar features are grouped into one cluster. Then the number of stations in each cluster is checked to see if there is any cluster containing very few stations (we used 10 as threshold in this analysis). A threshold of 10 is decided after discussions with NYCT staff. Clusters with fewer than 10 stations are deleted and their members became leftover ones. Clusters with more than 10 stations are kept. FASTCLUS will re-cluster those leftover ones to existing clusters. This procedure repeats until all clusters have more than ten stations.

In addition to cluster analysis with nineteen variables, we also conducted cluster analysis with principal components. This is done in two stages. In the first stage, we conducted a principal component analysis, followed by a cluster analysis with principal components in the second stage. Principal component analysis is a variable reduction procedure. When we have a large number of variables, it is possible that there is some redundancy in those variables. In this case, redundancy means that some of the variables are correlated with each other, possibly because they are measuring the same feature of time of day ridership pattern. Thus it may be possible to reduce these nineteen variables into a smaller number of principal components (artificial variables) that will account for most of the variance in the nineteen variables.

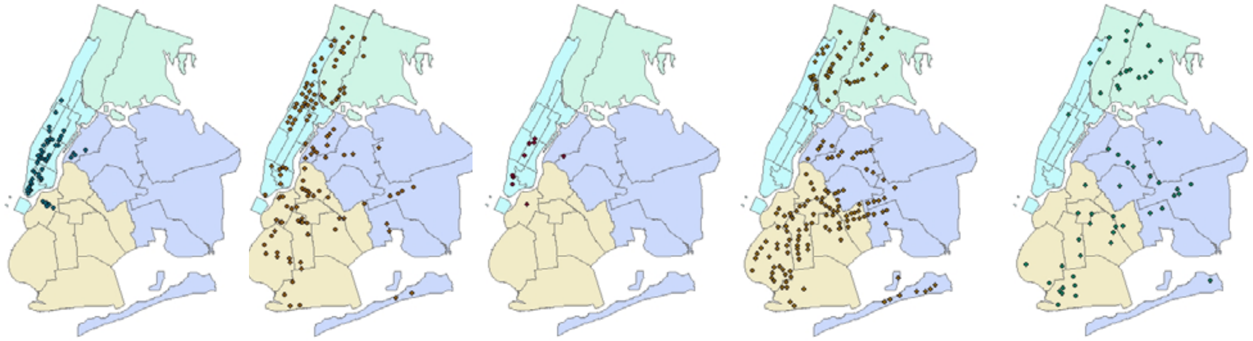
For our study, three principal components are identified: ridership ratio related component (e.g., morning peak over total daily ridership), transfer ridership ratio related component (e.g., the ratio of transfer ridership over total daily ridership), and ridership related component (e.g., morning peak ridership). A few variables (peak starting time and peak duration) can not be reduced into a principal component. Therefore, they remain as individual variables in the second stage, cluster analysis.

We compared the results of these two techniques (one is direct cluster analysis on the nineteen variables and the other is a principal component analysis followed by a cluster analysis). The first method, direct cluster analysis with nineteen variables, conveys a clearer time of day ridership pattern than the second technique, which includes a principal component analysis. We thus choose to adopt the direct cluster analysis method for this study.

Cluster Analysis Results

According to cluster analysis, five clusters are identified (see Figure 2-15).

Figure 2-15 Locations of Stations in Each Cluster

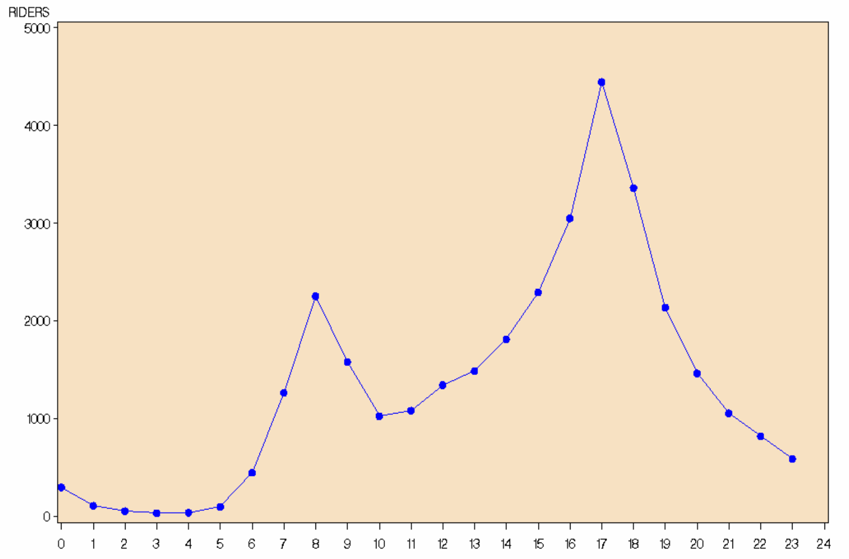


Cluster 1 Cluster 2 Cluster 3 Cluster 4 Cluster 5

Cluster 1

There are 64 stations in cluster 1. Stations in cluster 1 mostly locate in Manhattan CBD and core of Brooklyn and Queen (see Figure 2-15). On average, stations in this cluster have the highest total daily ridership, highest morning peak volume, highest afternoon peak volume, highest early morning hourly ridership, and highest midday hourly ridership. The typical feature of this pattern is that morning peak hourly ridership is relatively low while afternoon peak hourly ridership is relatively high. On average, morning peak starts at 8:00 a.m. and afternoon peak starts at 5:00 p.m. The average early morning hourly ridership is 315. This ridership climbs up sharply to 2,309 riders per hour in morning peak. During midday, ridership drops to 1,707 riders per hour. Then, it climbs up again to 4,514 riders per hour at around 5:00 pm (see Figure 2-16).

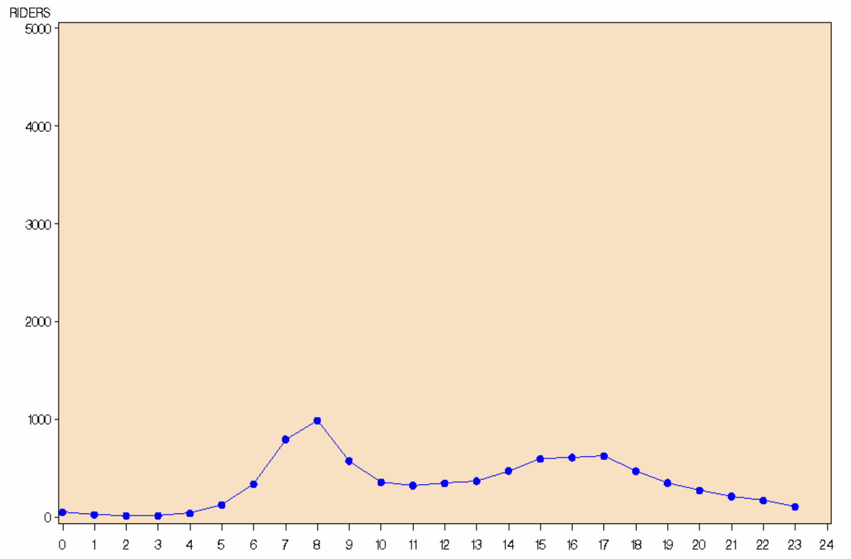
Figure 2-16 Ridership Temporal Distribution for Cluster 1



Cluster 2

There are 123 stations in cluster 2. Stations in cluster 2 mostly locate in upper Manhattan, Bronx, Brooklyn and Queens (see Figure 2-15). On average, stations in this cluster have the second lowest total daily ridership. Compared to cluster 1, afternoon peak hourly ridership is relatively low while morning hourly ridership is relatively high in cluster 2. Here "low" or "high" are based on the comparison between morning and afternoon peak hourly riderships for the same station. The absolute values of morning peak ridership and afternoon peak ridership are lower than those of cluster 1. Morning peak starts around 7:40 a.m. and afternoon peak starts around 4:10 p.m., both of which are earlier than that of cluster 1. The average early morning hourly ridership is 143. It then reaches 1,068 riders per hour in the morning peak. During midday, ridership drops to 707 riders per hour. Then it increases to 1,068 riders per hour in the afternoon peak (see Figure 2-17).

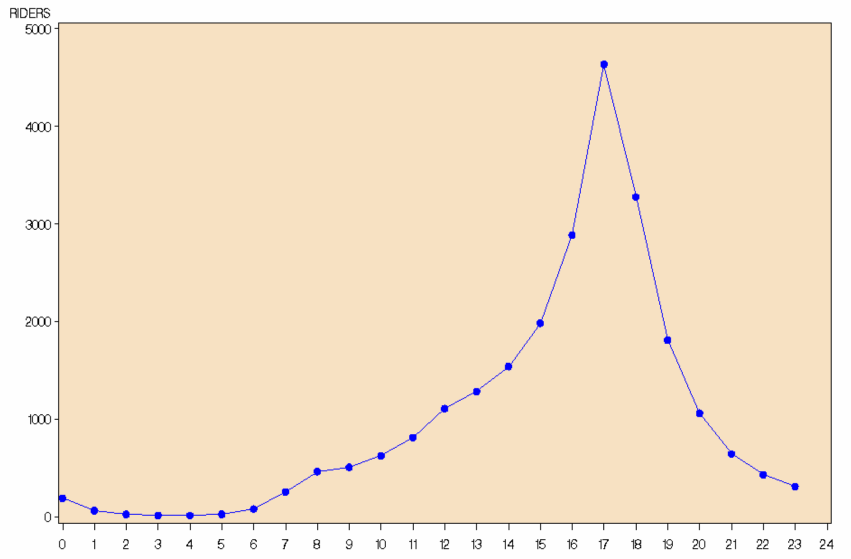
Figure 2-17 Ridership Temporal Distribution for Cluster 2



Cluster 3

There are 12 stations in cluster 3. Stations in cluster 3 mostly locate in midtown Manhattan, and core of Queens and Brooklyn, where the share of commercial land use is high (see Figure 2-15). On average, stations in this cluster have the highest afternoon peak hourly ridership in all clusters. A unique feature of this cluster is that there is only one afternoon peak during the entire day. The average early morning hourly ridership is 164. It then increases to 4,961 riders per hour, which is the afternoon peak (see Figure 2-18). In cluster 3, the afternoon peak starts at 4:52 pm.

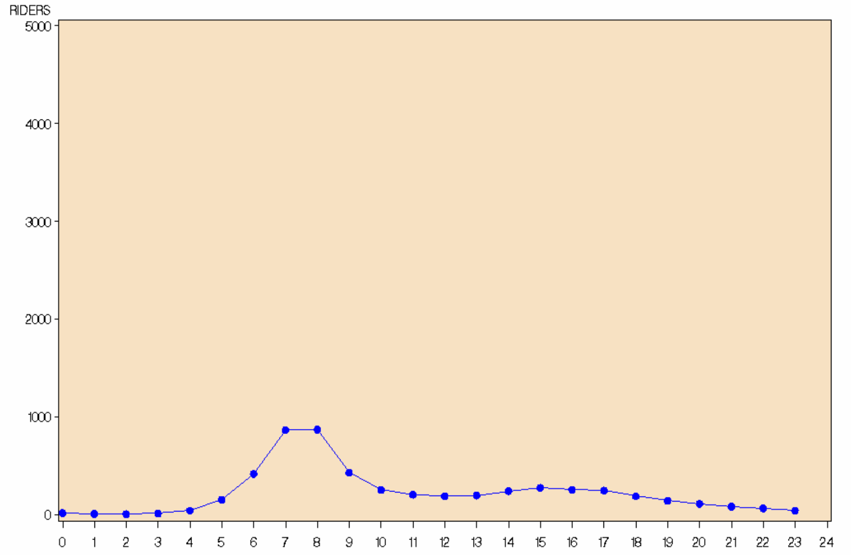
Figure 2-18 Ridership Temporal Distribution for Cluster 3



Cluster 4

There are 167 stations in cluster 4. Stations in cluster 4 mostly locate in uptown Manhattan, Bronx, Brooklyn and Queens, where the land use is mostly residential (see Figure 2-15). On average, stations in this cluster have the lowest total daily ridership, lowest morning peak volume, lowest afternoon peak volume, lowest early morning hourly ridership, and lowest midday hourly ridership. Compared to its afternoon peak volume, its morning peak volume is relatively high. The morning peak starts at 7:26 a.m., which is almost the earliest in all clusters (about the same as cluster 5). Afternoon peak starts at 3:39 p.m., which is the earliest in all clusters. Afternoon peak is quite insignificant in terms of its peak volume. The average midday hourly ridership is about 280, while the afternoon peak hourly ridership is only 297, which is slightly higher than the midday hourly ridership (see Figure 2-19).

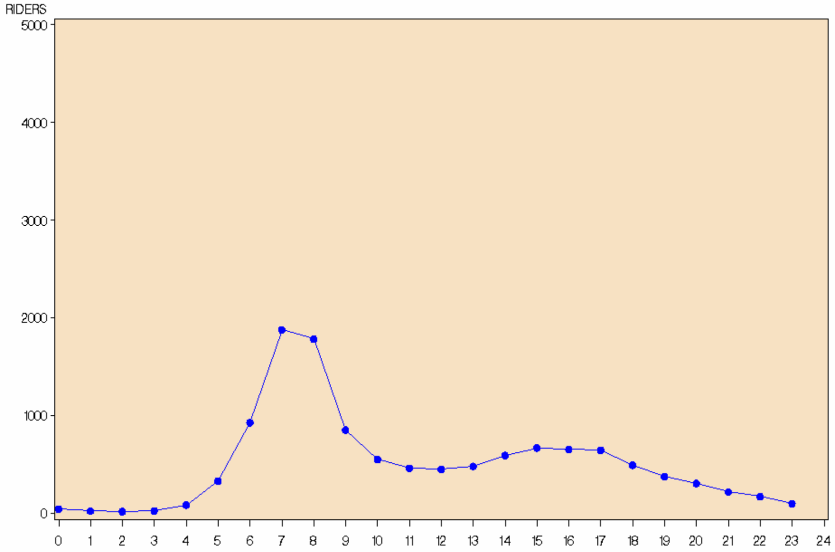
Figure 2-19 Ridership Temporal Distribution for Cluster 4



Cluster 5

There are 57 stations in cluster 5. Stations in cluster 5 mostly locate in areas surrounding terminal stations (see Figure 2-15). Therefore, one typical feature of this pattern is its high transfer ridership. Except that, its time of day pattern is quite similar to that of cluster 4. It has relatively high morning peak hourly ridership and relatively low afternoon peak ridership. Its morning peak starts at 7:26 a.m., which is almost the same as cluster 4. Its afternoon peak starts at 3:47 p.m., 8 minutes later than that of cluster 4. Because of transfer ridership, the hourly ridership in the morning peak, early morning, midday, and afternoon peak is larger than that of cluster 4. Ridership starts from 298 in early morning, followed by an increase to 2,113 in the morning peak period. The hourly ridership is 629 and 724 for midday and afternoon.

Figure 2-20 Ridership Temporal Distribution for Cluster 5



2-3. Summary

According to the cluster analysis, five clusters are identified based on nineteen variables discussed previously. The statistical means and standard deviations of each attribute for each of the five clusters are shown in Table 2-6.

Table 2-6 Means and Standard Deviations of the Characteristics of Subway Ridership's Pattern

Clusters	**Cluster 1**		**Cluster 2**		**Cluster 3**		**Cluster 4**		**Cluster 5**	
Number of stations	64		123		12		167		57	
Total Volume	mean	std	mean	std	mean	std	mean	std	mean	std
Total daily ridership	32,137	32,811	8,196	5,855	24,061	17,461	5,294	3,602	12,100	10,815
Concentration of the Volume										
Morning peak dummy	1	0	1	0	0	0	1	0	1	0
Afternoon peak dummy	1	0	1	0	1	0	1	0	1	0
Morning peak ridership volume	2,331	3,127	1,076	751	0	0	1,058	701	2,211	1,991
Morning peak hourly ridership	2,309	3,135	1,068	754	625	454	1,024	647	2,113	1,860
Morn. pk. hr. riders/total d. riders	0.06	0.03	0.13	0.03	0.03	0.01	0.20	0.03	0.18	0.03
Afternoon peak ridership volume	5,823	5,989	1,019	944	5,086	4,180	490	465	1,284	1,655
Afternoon peak hourly ridership	4,514	4,592	707	531	4,691	4,133	297	200	724	606
Early morning hourly ridership	315	456	143	93	164	129	137	86	298	270
Midday hourly ridership	1707	1635	447	318	1461	964	280	200	629	567
After. pk. hr. riders/total d. riders	0.15	0.05	0.09	0.02	0.18	0.05	0.06	0.02	0.06	0.02
Morn. pk. hr. riders/after. pk. hr. riders	0.49	0.41	1.66	0.59	0.18	0.09	3.78	1.29	3.00	1.16
E. morn. hr. riders/d. hr. riders	0.21	0.08	0.44	0.10	0.17	0.06	0.65	0.13	0.59	0.10
E. morn. hr. riders/mx. pk. hr. riders	0.06	0.03	0.13	0.03	0.04	0.02	0.14	0.03	0.14	0.03
Md. hr. riders/d. hr. riders	1.30	0.14	1.33	0.23	1.55	0.22	1.26	0.12	1.24	0.09
Md. hr. riders/mx. pk. hr. riders	0.37	0.09	0.41	0.10	0.39	0.14	0.27	0.05	0.31	0.08
Transfer Ridership										
M. pk. trans. riders/mx. pk. hr. riders	0.05	0.05	0.07	0.06	0.02	0.03	0.04	0.04	0.25	0.10
A. pk. trans. riders/mx. pk. hr. riders	0.02	0.03	0.06	0.05	0.01	0.01	0.05	0.05	0.24	0.08
E. morn. trans. riders/e. morn. hr. riders	0.08	0.09	0.100	0.09	0.05	0.07	0.05	0.05	0.38	0.17
Md. trans. riders/e. morn. hr. riders	0.03	0.03	0.07	0.06	0.02	0.02	0.04	0.04	0.25	0.10
Position of the Concentration										
M. pk. starting time	8.20	0.25	7.68	0.37	-	-	7.44	0.28	7.42	0.22
A. pk. starting time	16.89	0.40	16.17	0.76	16.86	0.25	15.65	0.74	15.79	0.74
M. peak duration	1.03	0.12	1.02	0.10	0.00	0.00	1.03	0.12	1.05	0.18
A. peak duration	1.28	0.40	1.37	0.59	1.13	0.23	1.61	0.73	1.61	0.67

3. Exploring the Relationships between the Built Environment and Time of Day Ridership Pattern – Why Do They Occur?

We will discuss what attributes are likely to affect time of day ridership pattern and how these attributes affect time of day ridership pattern.

3-1 Explanatory Attributes

We hypothesize that two categories of attributes will affect time of day ridership pattern. The first one relates to the local features of an area surrounding a station and the second set describes the relative position of a station in the study area. Variables in the first category include socio-demographic and economic variables and land use variables. These variables are calculated within a 500-meter radius of a station. The relative position of a station in the study area is measured by its generalized travel cost to CBD via subway. In our analysis, CBD is divided into three zones, midtown, downtown, and valley. Therefore there are three corresponding travel cost variables for each zone. The means and standard deviations of these variables for each cluster are shown in Table 3-1, 3-2, and 3-3.

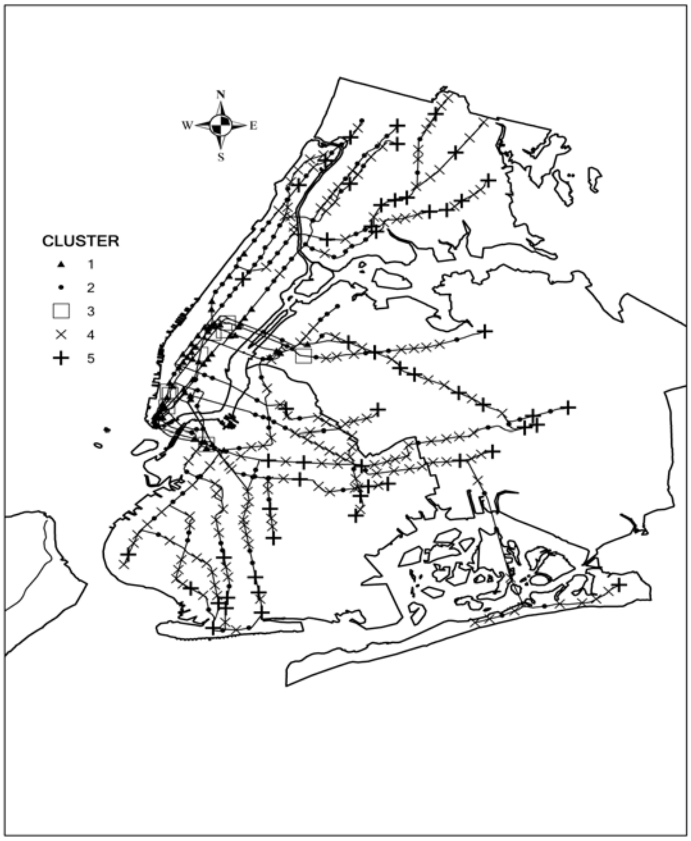
Stations in cluster 1 locate in areas with a high share of commercial land use (73.91%). On average, areas around these stations in cluster 1 have the highest values among all clusters in following attributes: number of households, number of white people, number of Asian, number of non-Hispanic population, number of people using public transit, number of people working at home, number of people using bicycle as a commute mode, number of people using walk as a commute mode, and number of workers. Office buildings take about 41.7% share in floor area for cluster 1, which is the highest among all land use types. This is followed by elevator apartments (17.84%). The average employment level in areas surrounding cluster 1 is 71,020, which is the second highest among all clusters. All these numbers indicate that the surrounding areas of cluster 1 stations could be described as a mix of commercial and residential zones, with an emphasis on commercial land use.

Stations in cluster 2 locate in areas with a mixed land use pattern (33.47% commercial and 66.53% residential). Residents in cluster 2 areas are most likely to use public transit among those in all five clusters. Population density is the highest in areas around cluster 2 stations. On average, areas around these stations have the highest among all clusters in following attributes: Hispanic population, number of elderly people, and factory land use. In cluster 2, the largest share of land use floor area is elevator apartments and walk up apartments, which are 34.32% and 21.99% respectively. The areas surrounding cluster 2 stations can be described as a mix of residential and commercial land use, but with an emphasis on residential land use.

Stations in cluster 3 locate in highly commercial areas. In areas surrounding cluster 3, commercial land use accounts for 86.82% of all floor area. People living in those areas are more likely to be white, work at home or walk to work locations, and have a higher income than residents in other clusters. The average employment density is the highest. About 52.95% of land use floor areas in cluster 3 are office buildings.

Stations in cluster 4 and cluster 5 share similar features. They locate in areas with a large share of residential land use (81.16% for cluster 4 and 73.40% for cluster 5). People living in those areas are more likely to be Africa American. The percentages of workers driving to work are relatively high in these clusters (26.5% for cluster 4 and 28.7% for cluster 5). The primary land use in cluster 4 and 5 is residential. Elevator apartments, walk up apartments, two family dwellings, and one family dwellings make up the largest floor areas, which are 31.23%, 22.03%, 11.11%, and 5.27%.

Figure 3-1: NYC Subway Stations by Cluster Membership



Geographically, there is a clear boundary to demarcate the clusters into two groups: clusters 1 and 3, and clusters 2, 4, and 5 (see Figure 3-1). The mean percentages of commercial floor areas in cluster 1 and cluster 3 are 73.91% and 86.82% respectively. The largest percentage in group 2 is cluster 2 (33.47%), which is less than half of the percentage for cluster 1. The total employment for clusters 1 and 3 are 71,020 and 86,172 respectively, followed by 6,026 for cluster 2, 2,553 for cluster 4, and 3,329 for cluster 5. Thus, the largest employment in group 2 is cluster 2, which is less than onetenth of the employment of cluster 1. Thus, the geographical separation between these two groups may be explained by the striking differences in the land use distribution of the surrounding areas.

Stations in clusters 1 and 3 are mingled together. And yet, they still depict a unique pattern. The twelve stations in cluster 3 seem to strategically locate in five positions: midtown, valley, downtown, core of Brooklyn, and core of Queens. Given that areas in cluster 3 are highly commercial, we might view areas in cluster 1 as five neighborhoods, which may have developed over time around these five commercial cores. Alternatively, it may be a collaborative process in which the two co-develop together.

The interlock of stations in clusters 2, 4, and 5 displays a different and yet more complicated pattern, compared with the one in clusters 1 and 3. First, cluster 2 seems to serve as the intermediate area between the borders of a commercial area (clusters 1 and 3) and a residential area (clusters 4 and 5). Second, as the distance from Manhattan increases further, some of the cluster 2 stations do appear at intermittent locations in the outlying areas, where stations in clusters 4 and 5 are mostly concentrated. These cluster 2 stations probably serve as local centers for those outlying areas. Third, the locations of cluster 5 stations also appear to be strategic as compared to those cluster 4 stations. We observe that cluster 5 stations are most likely situated in a location next to a large land area that is not served by a subway line, for example, terminal stations. This is reasonable, as these stations naturally serve as the connecting points between subway and other modes of transportation, for example, bus services.

To summarize, if we order the clusters by the intensity of commercial land use surrounding the station, the ranking goes: cluster 3, cluster 1, cluster 2, and clusters 4 and 5. Consequently, we can name these clusters as: commercial (high employment and low population), highly mixed-use (high population and employment), moderately mixed-use (high population and medium level of employment), residential (high population and low employment), and residential with transfer (high population, low employment, and high transfers).

Relating back to the characteristics observed on the time of day pattern, total daily ridership appears to be a function of the sum of population and employment. This is consistent with our expectations. However, the highest sum does not necessarily produce the highest ridership. While cluster 3 (commercial) has a higher sum of population and employment than cluster 1 (highly mixed use), the ridership of cluster 1 exceeds that of cluster 3. This probably suggests that a more balanced area, or a highly mixed-use area characterized by both high population and high employment, will likely produce the highest ridership.

Transfer ridership can significantly increase a station’s total daily volume, as shown in cluster 5 (residential with transfer). Clusters 4 (residential) and 5 are both primarily residential areas, but the ridership volume of cluster 5 is higher than cluster 4. An important difference between these two clusters is the level of commercialization.

Cluster 5 has a level of commercialization of 27%, which cluster 4 only has 17%. These additional commercial establishments in cluster 5 probably contribute to the ridership.

Even though commercial area is associated with afternoon peak and residential area is related to morning peak, a pure residential or a commercial area will not provide the highest morning or afternoon peak volumes. Instead, a highly mixed area, like cluster 1, possesses both the highest morning peak and afternoon peak volumes.

As expected, a residential area (e.g., clusters 4 and 5) is likely to have the earliest morning peak starting time and the longest morning peak duration. Contrary to our expectations, a commercial area (cluster 3) has neither the earliest afternoon peak starting time, nor the longest afternoon peak duration. Cluster 3, in fact, has almost the latest afternoon peak starting time and the shortest afternoon peak duration. Instead, a residential area (clusters 4 and 5) has the earliest starting time and the longest afternoon peak duration. Clusters 4 and 5, though being located in a primarily residential area, may still be more mixed use than a typical suburban residential neighborhood. Furthermore, as compared to those establishments in clusters 3 (which mainly comprises large corporations), the businesses in clusters 4 and 5 are likely small neighborhood businesses serving the neighborhood, for example, grocery stores, bakeries, salons, cleaners etc. These businesses operate in different hours, contributing to the formation of an early starting time and a long duration for the afternoon peak period.

A pattern that is shared by all clusters regardless of their land use distribution is that early morning hourly volume is always less than the average daily hourly volume and midday hourly volume is always larger than the average daily hourly volume. This indicates that most of the human activities take place during the day, instead of in the wee hours of the morning.

Table 3-1 Mean and Standard Deviation of Socio-economics and Demographic Variables within 500-meter Radius of Station

CLUSTER		H. mixed-use		M. mixed use		Commercial		Residential		Res. w. transf.	
Number of stations		64		123		12		167		57	
		MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
population		15,516	9,602	17,562	9,419	7,512	2,940	16,260	7,464	13,684	6,583
household		8,515	5,699	6,817	4,292	3,936	1,568	5,693	2,512	4,868	2,307
Race (ALL)	White Alone	11,284	8,011	7,070	6,569	5,498	2,452	5,465	3,929	4,688	3,373
	Black or Africa America	926	1,128	4,446	5,064	543	554	4,871	5,345	3,795	4,394
	American Indian	40	31	106	101	16	11	102	114	91	100
	Asian Alone	2,220	3,913	1,405	2,627	979	873	1,365	1,775	1,566	2,143
	Native Hawaii	9	13	14	20	9	9	12	17	14	18
	Some other	580	525	3,552	4,040	221	169	3,435	3,765	2,690	3,543
	Two race or more	459	265	982	665	244	101	1,003	665	841	646
Race (Hispanic)	Non_Hispanic	13,947	8,834	11,005	7,362	6,814	2,980	9,863	4,557	8,601	4,365
	Hispanic	1,572	1,219	6569	6,596	695	299	6,390	6,462	5,084	5,698
White Alone (%)		72.7%	/	40.2%	/	73.2%	/	33.6%	/	34.3%	/
Black or Africa America (%)		6.0%	/	25.3%	/	7.2%	/	30.0%	/	27.7%	/
American Indian (%)		0.3%	/	0.6%	/	0.2%	/	0.6%	/	0.7%	/
Asian Alone (%)		14.3%	/	8.0%	/	13.0%	/	8.4%	/	11.4%	/
Native Hawaii (%)		0.1%	/	0.1%	/	0.1%	/	0.1%	/	0.1%	/
Some other (%)		3.7%	/	20.2%	/	2.9%	/	21.1%	/	19.7%	/
Two race or more (%)		3.0%	/	5.6%	/	3.2%	/	6.2%	/	6.1%	/
Race (hispanic)	Non_Hispanic	89.9%	/	62.6%	/	90.7%	/	60.7%	/	62.9%	/
	Hispanic	10.1%	/	37.4%	/	9.3%	/	39.3%	/	37.1%	/
Public transit		5.148	3625	4290	3,201	2,253	1,096	3,602	1,942	2,979	1,732
Public transit	Bus	517	597	623	496	128	83	548	350	552	357
	Subway railroad	4.146	2797	3503	2,628	1,867	1,003	2,985	1,752	2,370	1,463
Work at home		692	477	268	395	445	304	130	121	104	120
Bicycle		83	76	53	79	43	56	26	28	14	21

CLUSTER		**H. mixed-use**		**M. mixed use**		**Commercial**		**Residential**		**Res. w. transf.**	
Number of stations		64		123		12		167		57	
Walk		2.719	1,763	914	1076	1,329	696	469	301	384	269
Auto		729	535	1,213	620	353	122	1,524	582	1,402	540
Public transit		54.9%	/	63.7%	/	50.9%	/	62.6%	/	61.0%	
Public transit	Bus	5.5%	/	9.2%	/	2.9%	/	9.5%	/	11.3%	/
	Subway railroad	44.2%	/	52.0%	/	42.2%	/	51.9%	/	48.5%	/
Work at home		7.4%	/	4.0%	/	10.1%	/	2.3%	/	2.1%	/
Bicycle		0.9%	/	0.8%	/	1.0%	/	0.5%	/	0.3%	/
Walk		29.0%	/	13.6%	/	30.0%	/	8.2%	/	7.9%	/
Auto		7.8%	/	18.0%	/	8.0%	/	26.5%	/	28.7%	/
Number of worker		9.455	5877	6,788	4,826	4461	1,954	5,787	2,697	4,911	2,465
Average commute time(minutes)		25	4	38	6	24	5	43	4	43	5
Departure time(hour)		8.7	0.5	8.6	0.5	8.3	0.8	8.7	0.3	8.5	0.7
Variation of departure time(hour)		2.3	0.4	2.8	0.4	2.2	0.5	2.9	0.3	2.8	0.4
Elderly		1.065	944	1126	686	486	244	1008	534	936	475
Median income		68.359	20,981	34,452	16,502	70,314	28,270	32,369	10,411	32,872	10,896
Bus stop		43	14	27	10	49	25	21	10	26	10
Express bus stop		16	15	1	3	23	22	1	2	2	3
Travel cost to downtown		7.6	2.3	12.1	3.0	7.7	2.2	13.5	2.8	14.8	2.6
Travel cost to midtown		7.1	1.6	11.3	3.1	7.1	1.9	13.6	2.8	13.9	2.8
Travel cost to valley		7.2	1.4	11.6	2.9	7.1	1.3	13.4	2.6	14.4	2.3

Table 3-2 Mean and Standard Deviation of Employment within 500-meter Radius of Station

CLUSTER	H. mixed-use		M. mixed use		Commercial		Residential		Res. w. transf.	
Number of stations	64		123		12		167		57	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
Total	71,020	47,800	6,026	7,196	86,172	64,408	2,553	1,631	3,329	2,337
Management	9,594	7,580	532	919	12,192	11,063	178	130	224	181
farmers	6	7	1	2	3	4	0	2	1	3
Business & Financial	6,918	6,638	240	720	8,380	8,430	58	55	99	107
Computer & mathematical	4,171	4,095	146	630	4,121	3,300	23	29	28	31
Architecture & engineering	1,084	731	64	132	1,255	710	16	25	27	49
Community& social service	819	549	166	176	854	564	68	56	108	118
Education, training & library	1,616	1,171	477	466	1,647	750	259	180	257	183
Arts, design & entertainment	4,549	4,194	224	454	5,998	5,934	54	73	51	57
Food preparation and serving	2,394	1,766	239	341	2,890	2,644	105	81	136	143
Sales and related	8,325	6,704	534	782	9,889	9,088	274	184	381	371
Office and administrative	12,843	9,491	876	1375	15,389	10,952	334	251	501	424
Production	2,582	2,274	302	370	3,617	2,407	166	189	152	117

Table 3-3 Land Use Share within 500-meter Radius of Station

CLUSTER		H. mixed-use		M. mixed use		Commercial		Residential		Res. w. transf.	
Number of stations		64		123		12		167		57	
		MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
Land use	commercial (%)	73.91%	0.76	33.47%	0.42	86.82%	0.89	18.84%	0.24	26.60%	0.33
	Residential (%)	26.09%	0.24	66.53%	0.58	13.18%	0.11	81.16%	0.76	73.40%	0.67
	Office (%)	63.91%	0.56	25.53%	0.33	65.59%	0.61	20.26%	0.21	22.77%	0.25
	RETAIL (%)	13.13%	0.17	14.26%	0.07	9.86%	0.07	24.09%	0.13	26.88%	0.20
	Garage (%)	2.26%	0.02	5.91%	0.04	1.81%	0.01	6.98%	0.07	9.72%	0.11
	Storage (%)	7.29%	0.10	8.69%	0.10	6.32%	0.09	8.33%	0.13	5.76%	0.09
	Factory (%)	0.64%	0.02	8.85%	0.10	0.78%	0.02	12.45%	0.24	6.08%	0.11
	Other (%)	12.77%	0.14	36.76%	0.36	15.63%	0.20	27.89%	0.21	28.79%	0.24
Percentage share in terms of total floor areas (compared within same cluster)											
one family dwellings		0.31%	0.01	1.70%	0.02	0.13%	0.00	4.58%	0.04	5.27%	0.04
two family dwellings		0.30%	0.00	4.51%	0.04	0.09%	0.00	13.64%	0.09	11.11%	0.07
walk up apartments		3.76%	0.04	21.99%	0.11	1.14%	0.01	29.16%	0.19	22.03%	0.17
elevator apartments		17.84%	0.12	34.32%	0.23	7.71%	0.06	29.19%	0.34	31.23%	0.29
warehouses		0.50%	0.02	2.86%	0.04	0.77%	0.02	2.23%	0.05	1.68%	0.03
Factory and industrial		0.50%	0.02	2.86%	0.04	0.77%	0.02	2.23%	0.05	1.68%	0.03
garages and gasoline		0.71%	0.01	1.44%	0.01	0.73%	0.01	1.05%	0.01	1.75%	0.02
hotels		3.98%	0.05	1.01%	0.02	5.51%	0.07	0.09%	0.00	0.16%	0.01
hospitals and health		0.41%	0.01	3.80%	0.07	0.23%	0.00	1.41%	0.03	1.43%	0.02
theatres		0.67%	0.02	0.34%	0.01	0.13%	0.00	0.07%	0.00	0.06%	0.00
store buildings		2.57%	0.02	2.90%	0.02	2.87%	0.02	2.93%	0.02	6.74%	0.07
loft buildings		5.12%	0.08	1.01%	0.02	5.36%	0.08	0.18%	0.01	0.07%	0.00
churches, synagogues		0.54%	0.01	3.47%	0.11	0.37%	0.00	1.44%	0.01	1.48%	0.01
asylums and homes		0.15%	0.00	0.32%	0.00	0.01%	0.00	0.36%	0.01	0.30%	0.01
office building		41.70%	0.37	3.99%	0.09	52.95%	0.46	0.99%	0.02	3.15%	0.05
public assembly and culture		0.41%	0.00	0.54%	0.01	0.50%	0.01	0.42%	0.01	0.49%	0.01
outdoor recreation		0.10%	0.00	0.38%	0.01	0.00%	0.00	0.05%	0.00	0.08%	0.00

CLUSTER	H. mixed-use		M. mixed use		Commercial		Residential		Res. w. transf.	
condominiums	16.63%	0.18	3.32%	0.05	18.11%	0.20	1.25%	0.03	2.21%	0.05
residence –multiple use	0.83%	0.01	2.39%	0.02	0.74%	0.01	4.26%	0.04	3.18%	0.03
transportation facilities	0.12%	0.00	0.22%	0.01	0.00%	0.00	0.02%	0.00	0.04%	0.00
utility bureau properties	0.00%	0.00	0.00%	0.00	0.00%	0.00	0.00%	0.00	0.00%	0.00
vacant land	0.02%	0.00	0.12%	0.00	0.02%	0.00	0.12%	0.00	0.07%	0.00
educational structures	2.26%	0.02	5.52%	0.04	1.30%	0.02	3.89%	0.03	4.57%	0.05
selected government installations	0.14%	0.00	0.38%	0.01	0.14%	0.00	0.28%	0.01	0.28%	0.01
miscellaneous	0.45%	0.01	0.61%	0.02	0.43%	0.01	0.15%	0.00	0.93%	0.03
Percentage share in terms of total number of buildings (compared within same cluster)										
one family dwellings	4%	0.06	11%	0.16	2%	0.03	17%	0.21	22%	0.24
two family dwellings	4%	0.08	23%	0.25	2%	0.03	36%	0.31	34%	0.30
walk up apartments	25%	0.26	34%	0.23	9%	0.08	29%	0.24	22%	0.19
elevator apartments	12%	0.09	6%	0.05	9%	0.08	2%	0.03	3%	0.03
warehouses	1%	0.04	2%	0.04	1%	0.04	1%	0.02	1%	0.02
Factory and industrial	1%	0.04	2%	0.04	1%	0.04	1%	0.02	1%	0.02
garages and gasoline	2%	0.02	2%	0.02	3%	0.03	2%	0.02	2%	0.02
hotels	2%	0.02	0%	0.00	3%	0.03	0%	0.00	0%	0.00
hospitals and health	0%	0.01	1%	0.01	0%	0.00	0%	0.00	0%	0.00
theatres	1%	0.01	0%	0.00	0%	0.01	0%	0.00	0%	0.00
store buildings	11%	0.07	4%	0.04	16%	0.08	3%	0.02	5%	0.05
loft buildings	7%	0.09	0%	0.01	12%	0.16	0%	0.00	0%	0.00
churches, synagogues	2%	0.01	2%	0.01	1%	0.01	1%	0.01	1%	0.01
asylums and homes	0%	0.00	0%	0.00	0%	0.00	0%	0.00	0%	0.00
office building	11%	0.07	1%	0.01	20%	0.14	0%	0.01	1%	0.02
public assembly and culture	0%	0.00	0%	0.00	1%	0.01	0%	0.00	0%	0.01
outdoor recreation	0%	0.00	0%	0.00	0%	0.00	0%	0.00	0%	0.00
condominiums	7%	0.05	2%	0.02	9%	0.11	0%	0.01	0%	0.01
residence –multiple use	8%	0.06	7%	0.07	8%	0.08	7%	0.07	6%	0.07
transportation facilities	0%	0.00	0%	0.00	0%	0.00	0%	0.00	0%	0.00

CLUSTER	H. mixed-use		M. mixed use		Commercial		Residential		Res. w. transf.	
utility bureau properties	0%	0.00	0%	0.00	0%	0.00	0%	0.00	0%	0.00
vacant land	0%	0.00	0%	0.00	0%	0.00	0%	0.00	0%	0.00
educational structures	1%	0.02	1%	0.01	1%	0.01	0%	0.00	0%	0.00
selected government installations	0%	0.00	0%	0.00	0%	0.00	0%	0.00	0%	0.00
miscellaneous	0%	0.00	1%	0.02	1%	0.01	0%	0.00	0%	0.00

3-2 Discrete Choice Analysis

Model specification

The descriptive analysis in Section 3-1 suggests a strong connection between the built environment and time of day ridership pattern. We conducted a discrete choice analysis to identify this relationship. In this study, each unit of observation represents a station. The dependent variable is station membership, identified through cluster analysis described above. The independent variables include two kinds: socio-demographic and economic attributes describing the local features of the area surrounding a station and generalized cost variables measuring travel cost from a station to the CBD area. The independent variables in this study are all alternative-specific. An alternative-specific variable can not be estimated with a single variable showing up in all utility functions. At maximum, an alternative-specific variable can be included in (J-1) utility functions, where J is the total number of clusters considered in the analysis (five in our study). This leaves at least one utility function in which the value for that variable is automatically set to be zero. The alternative whose utility function is set to zero is called base alternative. In this study, cluster 5 is the base alternative. In variable interpretation, the estimate is related to the utility function of the corresponding alternative. Take the "population" variable as an example. It can be entered into at most four of the five utility functions. A significant and negative coefficient of this variable in the utility function for cluster 3 suggests a higher likelihood of a cluster 5 membership than a cluster 3 membership.

According to the above discussion, the model has the following form:

$$\begin{aligned}
U_{i1} &= \beta_{10} + \beta_{11}x_{i1} + \beta_{12}x_{i2} + \ldots + \beta_{1n}x_{in} + \varepsilon_{i1} = V_{i1} + \varepsilon_{i1} \\
U_{i2} &= \beta_{20} + \beta_{21}x_{i1} + \beta_{22}x_{i2} + \ldots + \beta_{2n}x_{in} + \varepsilon_{i2} = V_{i2} + \varepsilon_{i2} \\
U_{i3} &= \beta_{30} + \beta_{31}x_{i1} + \beta_{32}x_{i2} + \ldots + \beta_{3n}x_{in} + \varepsilon_{i3} = V_{i3} + \varepsilon_{i3} \\
U_{i4} &= \beta_{40} + \beta_{41}x_{i1} + \beta_{42}x_{i2} + \ldots + \beta_{4n}x_{in} + \varepsilon_{i4} = V_{i4} + \varepsilon_{i4} \\
U_{i5} &= \varepsilon_{i5} = V_{i5} + \varepsilon_{i5}
\end{aligned} \tag{1}$$

Where,

$U_{i,1}$, $U_{i,2}$, $U_{i,3}$, $U_{i,4}$, and $U_{i,5}$ are the total utilities of station's belonging to clusters 1, 2, 3, 4, and 5 respectively;

V_{i1}, V_{i2}, V_{i3}, V_{i4}, and V_{i5} are the respective systematic utilities; note that since cluster 5 is taken as base alternative, V_{i5} is set to zero;

β_{10}, β_{20}, β_{30}, and β_{40} are constant terms;

x_{in} is the *n*th independent variable of station *i*;

β_{1n}, β_{2n}, β_{3n}, and β_{4n} are the coefficients of *n*th independent variable x_{in} for cluster *i* (where *i* = 1, 2, 3, or 4);

ε_{i1}, ε_{i2}, ε_{i3}, ε_{i4}, and ε_{i5} are error terms.

Assume that error terms are independently extremely value distributed, the probability of station's belonging to cluster *k* ($k \in \{1,2,3,4,5\}$) is expressed as:

$$P(i,k) = \frac{\exp(V_{ik})}{\sum_{k} \exp(V_{ik})}, \tag{2}$$

where β_{1n}, β_{2n}, β_{3n}, and β_{4n} are estimated from the model. Their interpretation goes as follows. If its estimate is greater than zero, a higher value on the corresponding variable would indicate a higher probability of a station's belonging to the respective cluster compared to cluster 5. On the other hand, if the estimate is less than zero, a higher value on the corresponding variable would suggest a lower probability of a station's belonging to the respective cluster compared to cluster 5.

Results

As indicated in Section 3.1, the independent variables considered in this study include socio-demographic variables (medium income, population, race composition, etc), land use information (commercial land use and residential land use), employment information (employment density), and relative location of a station with respect to CBD (generalized travel cost to midtown, downtown and valley).

Clustering of certain population in specific areas is evident in the New York Metropolitan Region. Highly commercial or mixed-use areas are mostly located in the CBD area and therefore have a dense transportation network, in terms of the number of transit stops. These areas are often clustered with high income people. Areas that are at a distance from CBD are mostly residential areas with households of middle or lower level incomes. The correlation of these variables prevents us from including all of them in the model. The final model (Table 3-4) only includes a subset of all available variables. Compared to the constant-only model, the log-likelihood is improved by 45% from -613 to -337, suggesting a clear link between the built environment and time-of-day ridership pattern. Note that even in this subset of all available variables, only a handful are statistically significant at 5% level, even though the final model has improved greatly over the constant-only one.

Table 3-4: Multinomial Logit Model Results (base alternative = cluster 5)

	Cluster 1		**Cluster 2**		**Cluster 3**		**Cluster 4**	
	64		**123**		**12**		**167**	
	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.
Constant	$\beta_{cont,1}$		$\beta_{cont,2}$		$\beta_{cont,3}$		$\beta_{cont,4}$	
	-0.24	2.11	2.62*	1.01	-6.86	5.04	-2.35*	1.09
Population (standardized)	$\beta_{pop,1}$		$\beta_{pop,2}$		$\beta_{pop,3}$		$\beta_{pop,4}$	
	1.02	0.65	0.43	0.33	-0.61	1.20	0.44	0.31
Percentage of White alone	$\beta_{white,1}$		$\beta_{white,2}$		$\beta_{white,3}$		$\beta_{white,4}$	
	5.87*	2.35	1.79*	0.79	8.39	5.33	0.36	0.71

Percentage of Asian alone	$\beta_{asian,1}$		$\beta_{asian,2}$		$\beta_{asian,3}$		$\beta_{asian,4}$	
	-0.41	2.96	-4.74*	1.62	4.49	8.50	-2.88	1.52
Commercial area (standardized)	$\beta_{c,1}$		$\beta_{c,2}$		$\beta_{c,3}$		$\beta_{c,4}$	
	7.76*	3.39	4.91	3.23	7.79*	3.40	-8.38	3.74
Distance to midtown (standardized)	$\beta_{midtown,1}$		$\beta_{midtown,2}$		$\beta_{midtown,3}$		$\beta_{midtown,4}$	
	-1.76	0.89	-0.57	0.29	-2.62*	1.29	-0.40	-1.76
Percentage of residential area	$\beta_{presi,1}$		$\beta_{presi,2}$		$\beta_{presi,3}$		$\beta_{presi,4}$	
	-5.15	4.15	-0.43	2.04	-4.63	7.41	0.77	-5.15
L(**c**)	-613							
$L(\hat{\boldsymbol{\beta}})$	-337							
ρ^2	0.45							

*: variables significant at 5% significance level.

On the socio-economic variables, a station located in an area with a high percentage of white population will most likely fall into clusters 1 and 2; the least probable group is clusters 4 and 5. A high percentage of Asians will be least likely to associate with a cluster 2 membership. A station surrounded with bounty commercial area will most likely belong to clusters 1, 2, and 3 and least likely to cluster 4. On the other hand, a station located in a primarily residential area will be least likely in clusters 1, 2, and 3. The three distance variables (distance to midtown, distance to valley, and distance to downtown) are highly correlated, so we only use the distance to midtown. Areas with a short distance to midtown are most likely located in cluster 3. The results from the discrete choice analysis are all consistent with the descriptive results described earlier.

4. Forecasting Time of Day Ridership Pattern

In this Chapter, we demonstrate a procedure to predict the time of day ridership pattern of a station based on its daily total ridership. This involves two steps sequentially. In the first step, we use a set of independent variables (those independent variables are used in our discrete choice model) to forecast a station's cluster membership (which cluster it belongs to). In the second step, we use the representative time of day ridership pattern of the predicted cluster and total daily ridership to forecast the actual time of day ridership pattern for a station. We will illustrate with one example: Chamber Street station. In the following, we describe each step in detail.

4-1 Step 1: Forecasting Cluster Membership

Three explanatory variables - population, commercial area, and distance to midtown – are standardized variables when used to forecast the cluster membership of station (see Chapter 3). Thus, prior to step 1, these variables needs to be standardized. The standardization of these variables is achieved by the following equation: $X_{s\tan dardized} = \frac{X - mean(X)}{std(X)}$, where *X* refers to a particular variable of interest, mean(*X*) is the average of *X*, and *std*(*X*) is the standard deviation of *X*. The mean and standard deviation of each variable are provided in Table 4-1. The other three explanatory variables – percentage of White alone, percentage of Asian alone, and percentage of residential floor area – are not standardized. However, it should be noted that these three variables are expressed in the number not a percentage format in the model. For instance, if percentage of White alone is 45%, number 0.45 is used in the computation instead of 45. Population, percentage of White alone, and percentage of Asian alone are calculated based on 2000 census data. Commercial area and percentage of residential floor area are calculated based on 2006 PLUTO file. The distance to midtown is measured by the generalized travel cost to midtown provided by NYCT.

Table 4-1 Mean and Standard Deviation of Explanatory Variables

Variables (units)	**Mean**	**Standard Deviation**
Population (persons)	15,995	8,395
Commercial area (sq. feet)	6,989,342	13,864,105
Generalized travel cost to midtown (dollars)	11.70	3.70

All variables except distance to midtown (provided by NYCT) are measured within the 500-meter radius of a station. For example, within the 500-meter radius around a station, 40% of area belongs to census tract A and 60% belongs to census tract B. The population within the 500-meter area around the station can be calculated as $\frac{\pi \bullet 500^2 \bullet 40\%}{Area(A)} \bullet Population(A) + \frac{\pi \bullet 500^2 \bullet 60\%}{Area(B)} \bullet Population(B)$,where $Area(A)$ and $Area(B)$ are area in square meters for census tracts A and B respectively.

Commercial area is measured as the total commercial floor area within the 500-meter area of a station. The percentage of residential floor area is calculated as the ratio of total residential floor area within 500-meter radius divided by the sum of residential and commercial floor area.

With these standardized independent variables, we can now forecast the probability of a station's belonging to a cluster using the following equations:

$$P_{i,1} = \frac{\exp(\beta_{cont,1} + \beta_{pop,1} \bullet X_{pop,s\tan dard,i} + \beta_{white,1} \bullet X_{white,s\tan dard,i} + \beta_{asian,1} \bullet X_{asian,s\tan dard,i}}{\sum\limits_{j=1,2,3,4} \exp(\beta_{cont,j} + \beta_{pop,j} \bullet X_{pop,s\tan dard,i} + \beta_{white,j} \bullet X_{white,s\tan dard,i} + \beta_{asian,j} \bullet X_{asian,s\tan dard,i}}$$

$$\frac{+\beta_{c,1} \bullet X_{c,s\tan dard,i} + \beta_{midtown,1} \bullet X_{midtwon,s\tan dard,i} + \beta_{presi,1} \bullet X_{presi,s\tan dard,i})}{+\beta_{c,j} \bullet X_{c,s\tan dard,i} + \beta_{midtown,j} \bullet X_{midtwon,s\tan dard,i} + \beta_{presi,j} \bullet X_{presi,s\tan dard,i}) + 1}. \tag{1}$$

$$P_{i,2} = \frac{\exp(\beta_{cont,2} + \beta_{pop,2} \bullet X_{pop,s\tan dard,i} + \beta_{white,2} \bullet X_{white,s\tan dard,i} + \beta_{asian,2} \bullet X_{asian,s\tan dard,i}}{\sum\limits_{j=1,2,3,4} \exp(\beta_{cont,j} + \beta_{pop,j} \bullet X_{pop,s\tan dard,i} + \beta_{white,j} \bullet X_{white,s\tan dard,i} + \beta_{asian,j} \bullet X_{asian,s\tan dard,i}}$$

$$\frac{+\beta_{c,2} \bullet X_{c,s\tan dard,i} + \beta_{midtown,2} \bullet X_{midtwon,s\tan dard,i} + \beta_{presi,2} \bullet X_{presi,s\tan dard,i})}{+\beta_{c,j} \bullet X_{c,s\tan dard,i} + \beta_{midtown,j} \bullet X_{midtwon,s\tan dard,i} + \beta_{presi,j} \bullet X_{presi,s\tan dard,i}) + 1}. \tag{2}$$

$$P_{i,3} = \frac{\exp(\beta_{cont,3} + \beta_{pop,3} \bullet X_{pop,s\tan dard,i} + \beta_{white,3} \bullet X_{white,s\tan dard,i} + \beta_{asian,3} \bullet X_{asian,s\tan dard,i}}{\sum\limits_{j=1,2,3,4} \exp(\beta_{cont,j} + \beta_{pop,j} \bullet X_{pop,s\tan dard,i} + \beta_{white,j} \bullet X_{white,s\tan dard,i} + \beta_{asian,j} \bullet X_{asian,s\tan dard,i}}$$

$$\frac{+\beta_{c,3} \bullet X_{c,s\tan dard,i} + \beta_{midtown,3} \bullet X_{midtwon,s\tan dard,i} + \beta_{presi,3} \bullet X_{presi,s\tan dard,i})}{+\beta_{c,j} \bullet X_{c,s\tan dard,i} + \beta_{midtown,j} \bullet X_{midtwon,s\tan dard,i} + \beta_{presi,j} \bullet X_{presi,s\tan dard,i}) + 1}. \tag{3}$$

$$P_{i,4} = \frac{\exp(\beta_{cont,4} + \beta_{pop,4} \bullet X_{pop,s\tan dard,i} + \beta_{white,4} \bullet X_{white,s\tan dard,i} + \beta_{asian,4} \bullet X_{asian,s\tan dard,i}}{\sum\limits_{j=1,2,3,4} \exp(\beta_{cont,j} + \beta_{pop,j} \bullet X_{pop,s\tan dard,i} + \beta_{white,j} \bullet X_{white,s\tan dard,i} + \beta_{asian,j} \bullet X_{asian,s\tan dard,i}}$$

$$\frac{+\beta_{c,4} \bullet X_{c,s\tan dard,i} + \beta_{midtown,4} \bullet X_{midtwon,s\tan dard,i} + \beta_{presi,4} \bullet X_{presi,s\tan dard,i})}{+\beta_{c,j} \bullet X_{c,s\tan dard,i} + \beta_{midtown,j} \bullet X_{midtwon,s\tan dard,i} + \beta_{presi,j} \bullet X_{presi,s\tan dard,i}) + 1}. \tag{4}$$

$$P_{i,5} = \frac{1}{\sum\limits_{j=1,2,3,4} \exp(\beta_{cont,j} + \beta_{pop,j} \bullet X_{pop,s\tan dard,i} + \beta_{white,j} \bullet X_{white,s\tan dard,i} + \beta_{asian,j} \bullet X_{asian,s\tan dard,i}}$$

$$\frac{}{+\beta_{c,j} \bullet X_{c,s\tan dard,i} + \beta_{midtown,j} \bullet X_{midtwon,s\tan dard,i} + \beta_{presi,j} \bullet X_{presi,s\tan dard,i}) + 1}. \tag{5}$$

Where,

$P_{i,1}$, $P_{i,2}$, $P_{i,3}$, $P_{i,4}$ and $P_{i,5}$ are the probabilities of station *i*'s belonging to clusters 1, 2, 3, 4 and 5 respectively;

$\beta_{cont,1}$, $\beta_{cont,2}$, $\beta_{cont,3}$, and $\beta_{cont,4}$ [1] are constants for clusters 1, 2, 3 and 4 respectively;

[1] Note that all coefficients for cluster 5 are set to zero because cluster 5 is the base alternative.

$\beta_{pop,1}$, $\beta_{pop,2}$, $\beta_{pop,3}$, and $\beta_{pop,4}$ are coefficients of population for clusters 1, 2, 3, and 4 respectively;
$\beta_{white,1}$, $\beta_{white,2}$, $\beta_{white,3}$, and $\beta_{white,4}$ are coefficients of the percentage of white population for clusters 1, 2, 3, and 4 respectively;
$\beta_{asian,1}$, $\beta_{asian,2}$, $\beta_{asian,3}$, and $\beta_{asian,4}$ are coefficients of the percentage of Asian population for clusters 1, 2, 3, and 4 respectively;
$\beta_{c,1}$, $\beta_{c,2}$, $\beta_{c,3}$, and $\beta_{c,4}$ are coefficients of commercial floor area for clusters 1, 2, 3, and 4 respectively;
$\beta_{midtown,1}$, $\beta_{midtown,2}$, $\beta_{midtown,3}$ and $\beta_{midtown,4}$ are coefficients of generalized travel cost to midtown for clusters 1, 2, 3, and 4 respectively;
$\beta_{presi,1}$, $\beta_{presi,2}$, $\beta_{presi,3}$ and $\beta_{presi,4}$ are coefficients of the percentage of residential floor area for clusters 1, 2, 3, and 4 respectively;
$X_{pop,s\tan dard,i}$ is the standardized population in station *i*;
$X_{white,s\tan dard,i}$ is the percentage of white population in station *i*;
$X_{asian,s\tan dard,i}$ is the percentage of Asian population in station *i*;
$X_{c,s\tan dard,i}$ is the standardized commercial floor areas in station *i*;
$X_{midtwon,s\tan dard,i}$ is the standardized generalized travel cost to midtown from station *i*;
$X_{presi,s\tan dard,i}$ is the percentage of residential floor areas in station *i*.

Using the estimated coefficients shown in Table 3-4 (in Chapter 3), we can now calculate the probability of a station's belonging to a cluster as follows. Let us assume a station has the following attributes: 15,000 residents, 9,762,000 square feet of commercial floor area; the percentage of residential land use is 0.45; the percentage of White alone is 0.75; and the share of Asians is 0.05. The generalized travel cost of this station to midtown is 12. After standardization (using the mean and standard deviation of the existing dataset), these variables become: 0.1185 for population, 0.2 for commercial floor area, and 0.0810 for the generalized travel cost to midtown. The other variables remain unchanged, as they do not need to be standardized. With these numbers, we can then calculate the probability of this station's belonging to cluster 1, cluster 2, cluster 3, cluster 4 and cluster 5. They are: 2%, 60%, 2%, 8%, and 27% respectively. In other words, this station most likely belongs to cluster 2.

4-2 Step 2: Forecasting Time of Day Pattern

In this section, we describe how we will forecast the time of day ridership pattern for a station. To achieve this, our first step is to forecast the cluster membership of the station using the procedure described in Section 4-1. In addition, we need three additional pieces of information: average daily ridership of the most likely cluster, average hourly ridership of the most likely cluster, and total daily ridership of the station of interest. Information on the first two variables are readily available from our analysis. Figures 2-16 to 2-20 in Chapter 2 graphically depict the average hourly ridership for each cluster. We now present them in a table format in Table 4-2.

Table 4-2 Average Hourly Ridership and Daily Ridership for the Five Clusters

Time of day	Hourly Ridership				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
	H. mixed-use	**M. mixed use**	**Commercial**	**Residential**	**Res. w. transf.**
12:01-1:00am	297	50	191	16	43
1:01-2:00am	108	23	63	8	22
2:01-3:00am	52	14	26	6	16
3:01-4:00am	32	15	15	12	24
4:01-5:00am	37	39	13	41	83
5:01-6:00am	98	125	26	154	330
6:01-7:00am	447	336	80	417	926
7:01-8:00am	1,265	796	256	866	1,877
8:01-9:00am	2,254	988	464	871	1,783
9:01-10:00am	1,580	575	507	428	851
10:01-11:00am	1,026	358	625	253	549
11:01-12:00pm	1,082	323	812	205	460
12:01pm-1:00pm	1,344	345	1,110	188	448
1:01-2:00pm	1,486	369	1,285	196	479
2:01-3:00pm	1,812	473	1,541	241	591
3:01-4:00pm	2,293	595	1,985	275	668
4:01-5:00pm	3,051	611	2,889	255	654
5:01-6:00pm	4,448	623	4,637	247	645
6:01-7:00pm	3,363	470	3,278	189	491
7:01-8:00pm	2,137	350	1,811	141	374
8:01-9:00pm	1,463	274	1,063	110	303
9:01-10:00pm	1,054	212	644	82	217
10:01-11:00pm	818	170	431	64	169
11:01-12:00am	588	106	309	38	97
Average Daily Total	32,137	8,240	24,061	5,302	12,100

The last piece of information: total daily ridership of the station of interest, may be obtained from various sources: from historical data, or from a predicted ridership model.

The central idea of the forecasting process is to scale the average hourly ridership during different hours of the most likely cluster to a particular station by applying a scale factor. This scale factor is calculated as the ratio of dividing total daily ridership of a station by the average daily total ridership of the most likely cluster. We now illustrate this process with an example: Chamber Street station in lower Manhattan.

Example: Forecasting Time of Day Ridership for Chamber Street Station

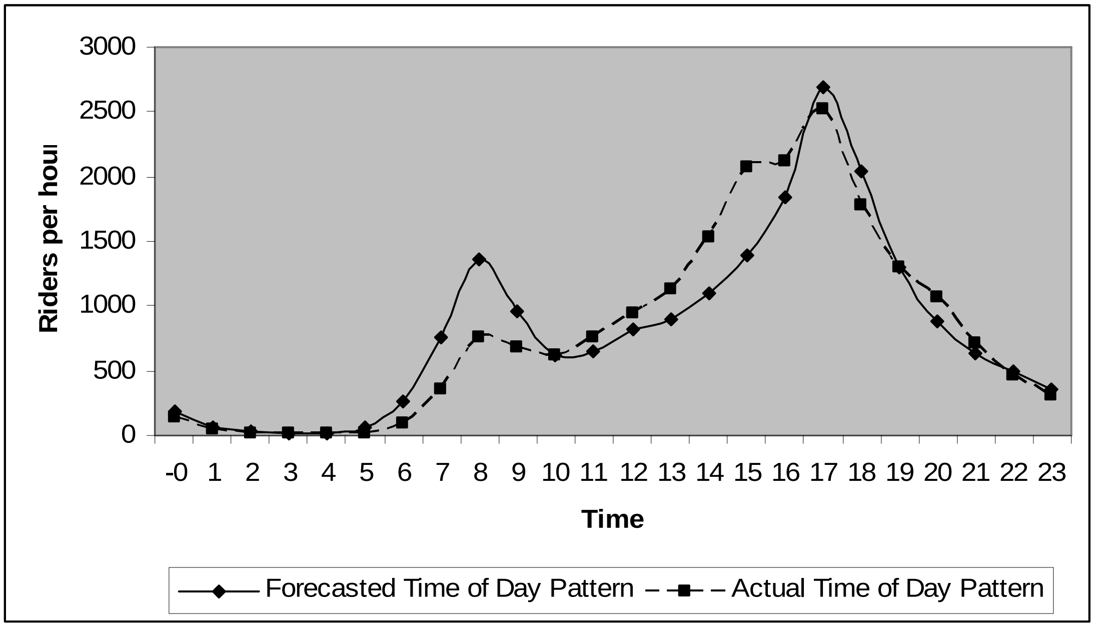
Step 1: Forecast the cluster membership of the Chamber Street Station

We first standardize the six independent variables for the Chamber Street Station. The standardized values for population, percentage of white population, percentage of Asian population, generalized cost to midtown, percentage of residential floor area, and commercial floor area are: -0.89, 0.78, 0.12, -1.11, 0.28, and 1.16 respectively. Using equations (1) to (5), the projected probabilities of the Chamber Street Station's belonging to clusters 1, 2, 3, 4, and 5 are: 0.8, 0.02, 0.18, 0, and 0 respectively. In this case, the Chamber Street Station most likely belongs to Cluster 1.

Step 2: Forecast time of day pattern

Given that the Chamber Street station most likely belongs to cluster 1, we can now calculate a scale factor for this particular station. This is obtained by dividing the total daily ridership of the Chamber Street station (which is 19,446) by the average daily ridership of cluster 1 (which is 32,127, see Table 4-1). The scale factor is equal to 0.61. Using this scale factor, we will adjust the average hourly ridership of cluster 1 for the Chamber Street station. To do this, we multiply the average hourly ridership of cluster 1 for each hour by 0.61. The result is the predicted hourly ridership for the Chamber Street station. Figure 4-1 graphically shows the forecasted and the actual time of day ridership pattern for the Chamber Street station.

Figure 4-1: Forecasted vs. Actual Time of Day Ridership Pattern for the Chamber Street Station



We applied this procedure to every station in New York City and created a forecasted time of day pattern for every station. The zip file TOD_Project in the enclosed CD contains all the information. The user can also use “time_of_day_forecast.sas” file in the sascode zip file to generate new forecasted time of day pattern.

5 Summary and Future Works

5-1 Summary

In this study, we focused on the time of day aspect of station ridership in New York City. During the process, we used the average subway ridership data every 6 minutes for one day in May 2005, the 2000 socio-economic and demographic data from the Census Bureau, employment data from the CTPP package, land use information from the PLUTO file released in 2006, and generalized travel cost information by station provided by NYC Transit.

Analysis on stations' time of day ridership pattern shows differences between weekdays and weekends. In our preliminary analysis on stations along the number 7 line, we observe four different ridership patterns: high morning peak pattern, high afternoon peak pattern, no morning peak pattern, and evening peak pattern. On Saturday, we observe two additional patterns in addition to the four observed on weekdays: a very long peak period (a peak starts in the morning and ends in the afternoon) and a morning peak only pattern. On Sunday, some of the patterns observed on Saturday disappeared.

When we analyze the weekday time of day ridership pattern for all stations in New York City, we identified five distinctive patterns. Geographically, these five clusters can be divided into two groups: clusters 1 and 3, and clusters 2, 4, and 5. Stations in cluster 1 and 3 more likely locate in Manhattan CBD or business centers of other boroughs. Therefore, the time of day patterns of clusters 1 and 3 have a very high afternoon peak volume, which are apparently the result of local commercial land use. In cluster 1, the morning peak volume is lower than the afternoon peak volume. In cluster 3, the morning peak even disappears. Stations in cluster 2, 4, and 5 more likely locate in uptown Manhattan, Brooklyn, Bronx, and Queens. Compared to stations in cluster 1 and 3, the commercial land use share is much lower while residential land use is higher in the stations of cluster 2, 4, and 5. Among cluster 2, 4, and 5, areas surrounding cluster 2 stations have a medium level in terms of the share of commercial land use and residential land use, which could be described as mixed land use pattern. Areas surrounding clusters 4 and 5 stations have a high share of residential land use. Correspondingly, we observed a higher ratio of morning peak hourly ridership over afternoon peak hourly ridership for clusters 4 and 5 stations than cluster 2 stations. The distinction between cluster 4 and cluster 5 comes from the transfer ridership. Stations in cluster 5 have much higher transfer ridership. On average, more than 24% of hourly ridership in cluster 5 stations is transfer ridership. This may result from the fact that cluster 5 stations most likely locate in a location next to a large land area that is not served by a subway line, for example, the terminal stations.

According to discrete choice analysis, both socio-economic and demographic features and land use features are statistically significant in terms of explaining the formation of time of day patterns, especially for commercial floor areas, the coefficients of which are significant for all clusters. The regression results could be applied in the projection of the time of day patterns of new constructed stations. The comparison of forecasted patterns and actual patterns in existing stations indicates that there is 76% of projection matches actual patterns.

5-2 Discussion

The linkage between land use and time of day pattern suggests that time is an important dimension in demand and supply analysis. Two stations can have similar total daily riderships, but possess distinctive time of day patterns. One example is Kew Gardens in Queens and 53rd street in Midtown, Manhattan. Both have a daily ridership of about 25,000. Kew Gardens has a morning peak volume of 4,651, while 53rd street has 0. Kew Gardens station has an afternoon peak that is less than one-third of that of 53rd street; its early morning ridership is more than 10 times that of 53rd street. Despite the small volume, the duration of the afternoon peak period for Kew Gardens is one and half times that of 53rd street. Kew Gardens station is located in a residential area with a large number of transfers and a membership in cluster 5, while 53rd street station is situated in a primarily commercial area with a membership in cluster 3.

The notion that the land use influences the time of day ridership pattern is important, especially amidst the recent surge of interest in the compact and mixed-use neighborhoods. Within the context of New York City, a rise in the level of commercialization will likely increase the afternoon peak volume, but decrease the morning peak volume. A further rise in commercialization might also make the morning peak period disappear and sharpen the afternoon peak period. Contrary to our hypothesis, a residential area tends to have longer morning and afternoon peak periods.

The study results can also have practical implications in the service planning of transit agencies. This may be reflected in several aspects. First, the current study develops a series of variables to classify time of day ridership pattern. This method can also be used to classify bus ridership during a day. Second, given the information on socio-economic, land use and transportation characteristics of the surrounding area, the current study provides a method to forecast the membership of time of day ridership pattern and then use that information to forecast the time of day ridership pattern for a particular station. Such information can be very useful in service scheduling at the station as well as complementary services provided by other modes, for example, bus service or parking garages nearby.

5-3 Future Works

Our study has confirmed the link between the land use and time of day patterns of weekdays. However, some other interesting questions still wait to be answered. The current results only apply to weekday station ridership based on data in May 2005. It would be worthwhile to investigate station ridership during other time frames. In addition, our preliminary analysis on stations along the number 7 line reveals quite different time of day pattern between weekdays and weekends. Thus, an analysis on the time of day pattern of weekend ridership is warranted.

A trip is a movement in space. In other words, the time of day pattern of subway ridership observed at the station level is likely not only a function of the local environment, but also the attributes of other locations. In the current study, we use the distance to the CBD area to approximate the relative location of a station in space, with respect to the CBD area. Future studies can examine the diurnal pattern of adjacent stations to understand how time of day ridership pattern might evolve between stations.

Appendix

The main purpose of this analysis is to understand the spatial distribution of total daily riderships in New York City. We seek to answer two main questions: 1) how does ridership distribution vary in space (across New York City)? and 2) what specific spatial patterns can we identify?

A. Ridership Spatial Distribution

In Phase 1, we investigate the spatial distribution of ridership. We hypothesize that the spatial distributions of total daily ridership are different for weekdays and weekends. The patterns for weekday, Saturday and Sunday are therefore separately analyzed.

A-1 Spatial Distribution of Weekday Ridership

The histogram of weekday total daily ridership is shown in Figure A-1. The X-axis represents the number of riders per day. The Y-axis represents the count - the number of stations that fall into a particular ridership range, such as 0-1,000, 1,001-2,000, 2,001-3,000, etc. The overall ridership pattern fits a log-normal distribution. Most of the stations' (about 95%) total daily riderships are under 20,000. The maximum daily ridership is at Times Square – 42nd Street station, which on average has 173,260 riders per day.

Figure A-1 Weekday Daily Ridership Distribution

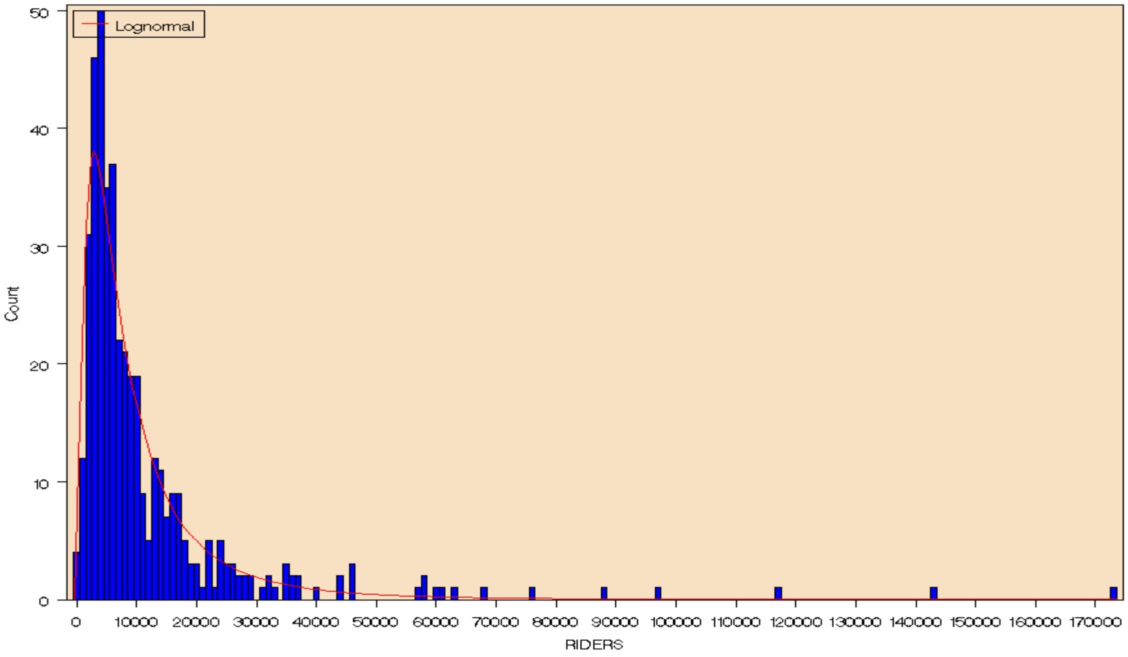


Table A-1 shows some descriptive statistics of the ridership information for weekday. On average, the mean total daily ridership is 11,649, and the standard deviation is 16,933. Seventy five percent of the stations have a total daily ridership under 13,112 and 25% of stations have a total daily ridership under 3,776. The skewness of ridership is 4.95, which indicates the distribution is screwed to the right, consistent with our observation in Figure A-1.

Table A-1 Statistical Description of Daily Ridership of New York City Subway

Statistic	Mean	Maximum	25 Percentile	50 Percentile	75 Percentile
Ridership	11,649	173,260	3,776	6,392	13,112

The spatial distribution of daily ridership is shown in Figure A-2. In general, most of stations with high riderships are located in Manhattan and Queens, while most of stations with low riderships are located in Bronx and Brooklyn.

Figure A-2 Weekday Spatial Ridership Distribution

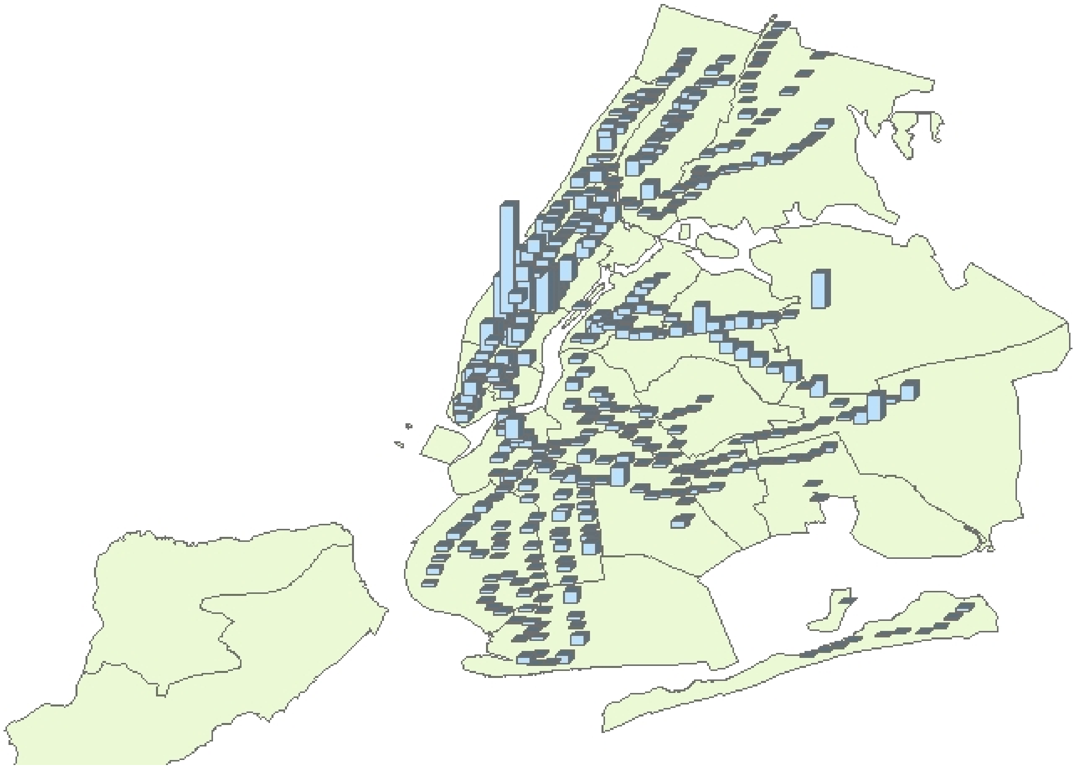


Figure A-2 presents the spatial distribution of station ridership for weekday. Figure A-3 shows the distribution of ridership at the borough level. In general, Manhattan has most number of stations (more than 50) whose riderships are above 75 percentile of all station riderships in the city while Bronx has the fewest stations (fewer than 10) whose riderships are above 75 percentile. Brooklyn has the most number of stations whose riderships are under 25 percentile (more than 50), while Manhattan has the fewest stations (fewer than 10) under 25 percentile. In Manhattan, Brooklyn and Bronx, more stations fall into 25 to 50 Percentile, while in Queens, more stations fall under 25 Percentile.

Figure A-3 Spatial Distribution of Subway Ridership in terms of the Number of Stations in each Borough that Falls in a Particular Percentile

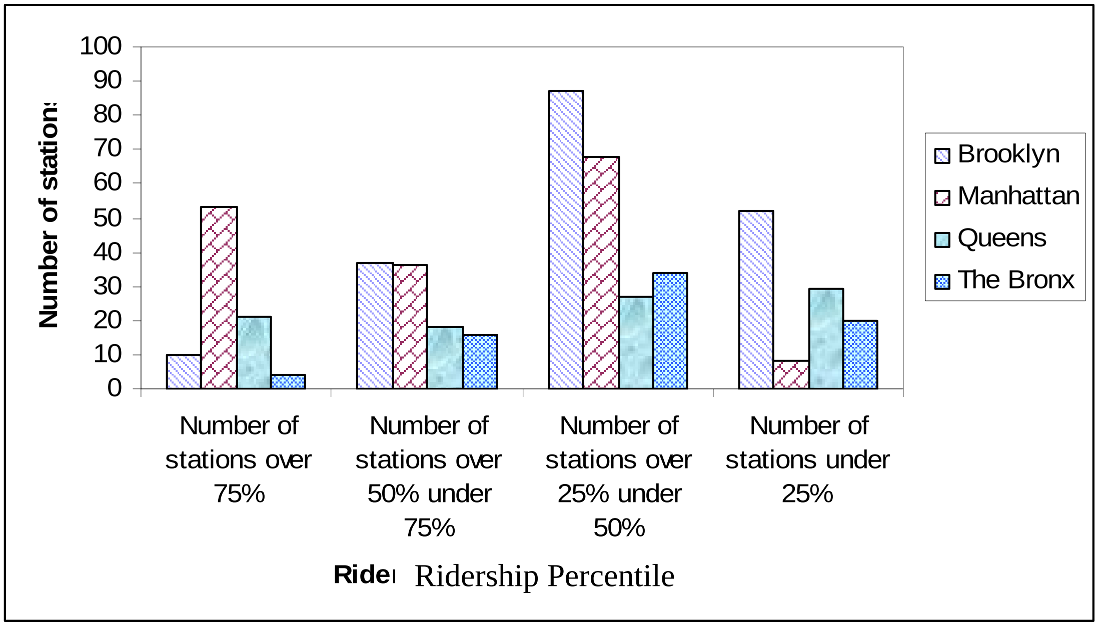
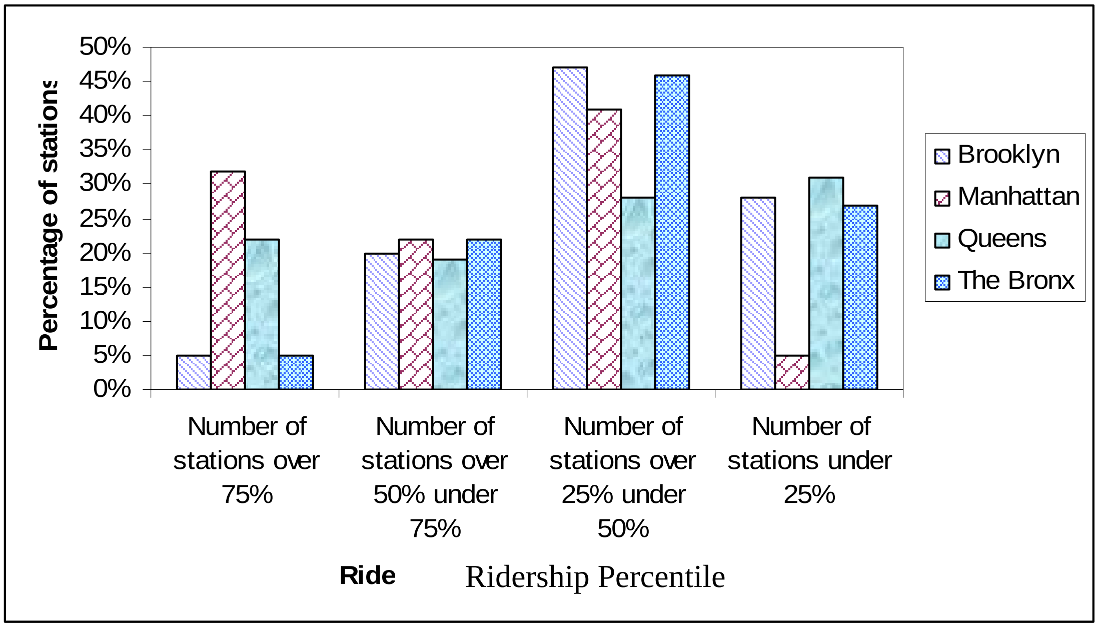


Figure A-3 shows the absolute number of stations in each percentile. Figure A-4 shows the percentage of stations in each borough whose ridership falls in a particular percentile. In sum, 32% of stations in Manhattan are above 75% percentile in ridership, which are highest in all four counties. Only 5% of stations in Manhattan are under 25% percentile in ridership. Brooklyn and Bronx both have lowest percentage (5%) of stations that are above 75% percentile in ridership. Queens has relatively higher percentage - about 22% - of stations which are above 75% Percentile in ridership. All three counties – Brooklyn, Bronx, and Queens – have about 30% of stations which are under 25% percentile in ridership. From the patterns observed, stations in Manhattan have the highest probability to have high ridership while stations in Brooklyn and Bronx have the lowest probability to have high ridership.

Figure A-4 Spatial Distribution of Station Ridership in terms of the Percentage of Stations in Each Borough that Falls in a Particular Percentile



A-2 Saturday

The histogram of Saturday total daily ridership is shown in Figure A-5. The X-axis and Y-axis have the same meanings as Figure A-1. On Saturday, most of stations' (about 95%) riderships fall under 14,000. The maximum daily ridership still occurs in Time square – 42nd Street station, which has 118,257 riders per day on average. The average total daily ridership in Saturday is only half of that in weekday, which may due to decrease in work trips. Other percentile information is shown in Table A-2. On Saturday, the mean total daily ridership is 6,494, with standard deviation to be 8,537. The 75% percentile of ridership is 7,615 and 25% percentile is 1,951. Compared to weekday, the average daily ridership in Saturday reduced by 44 percents, and the maximum ridership reduced by 32 percents. An interesting pattern is that the ratios of Saturday statistics over weekday (mean, 25%, 50% and 75% percentile) fall into a narrow range (0.51~0.58).

Figure A-5 Saturday Daily Ridership Distribution

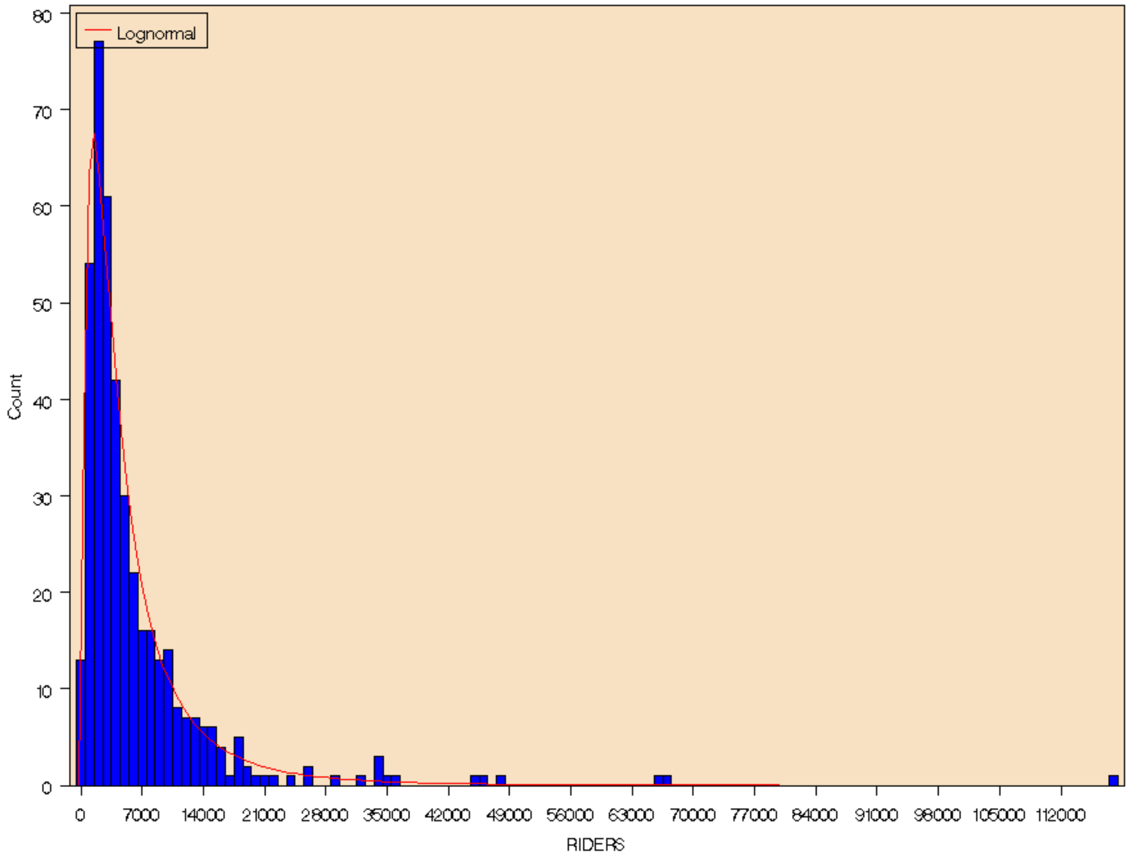
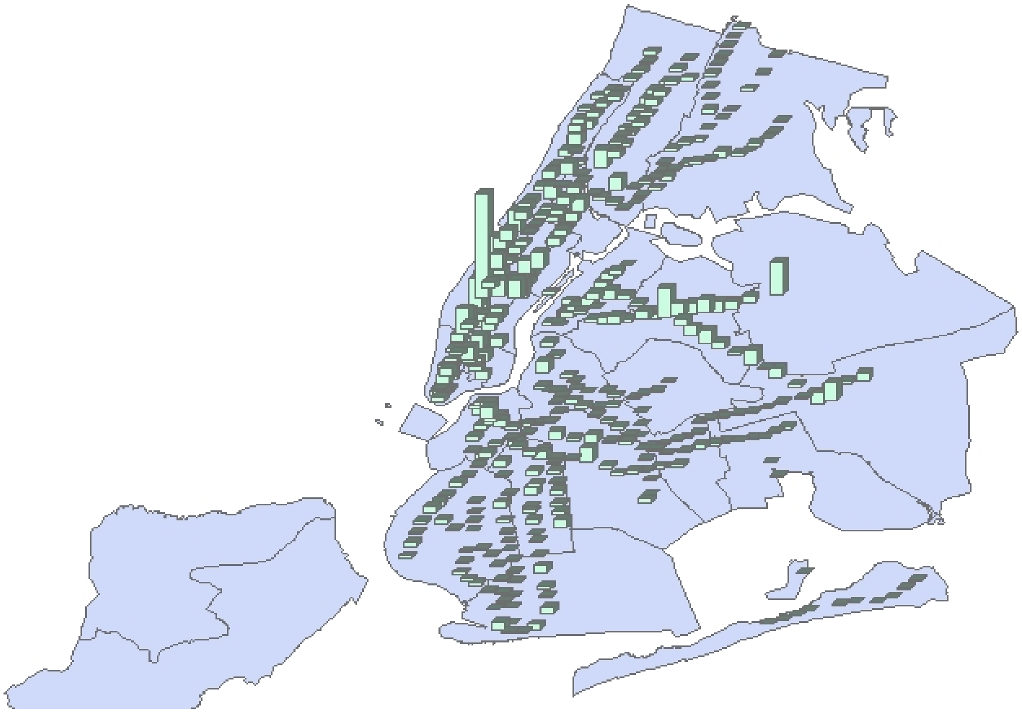


Table A-2 Statistical Description of Saturday Daily Ridership

Statistic	Mean	Maximum	25 Percentile	50 Percentile	75 Percentile
Ridership	6,494	118,257	1,951	3,700	7,615

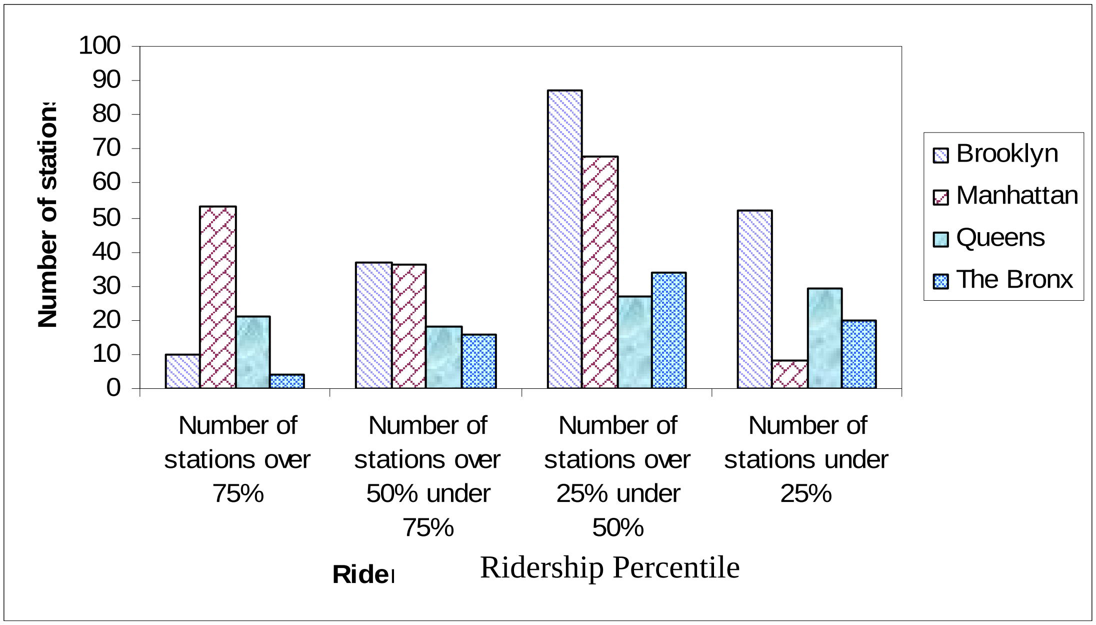
The spatial distribution of daily ridership is shown in Figure A-6 at city level. The Saturday ridership spatial distribution is similar to weekday pattern. Most of stations with higher ridership are still located in Manhattan and Queens, while most of stations with lower ridership are still located in Bronx and Brooklyn.

Figure A-6 Spatial Distribution of Subway Ridership for Saturday by Station



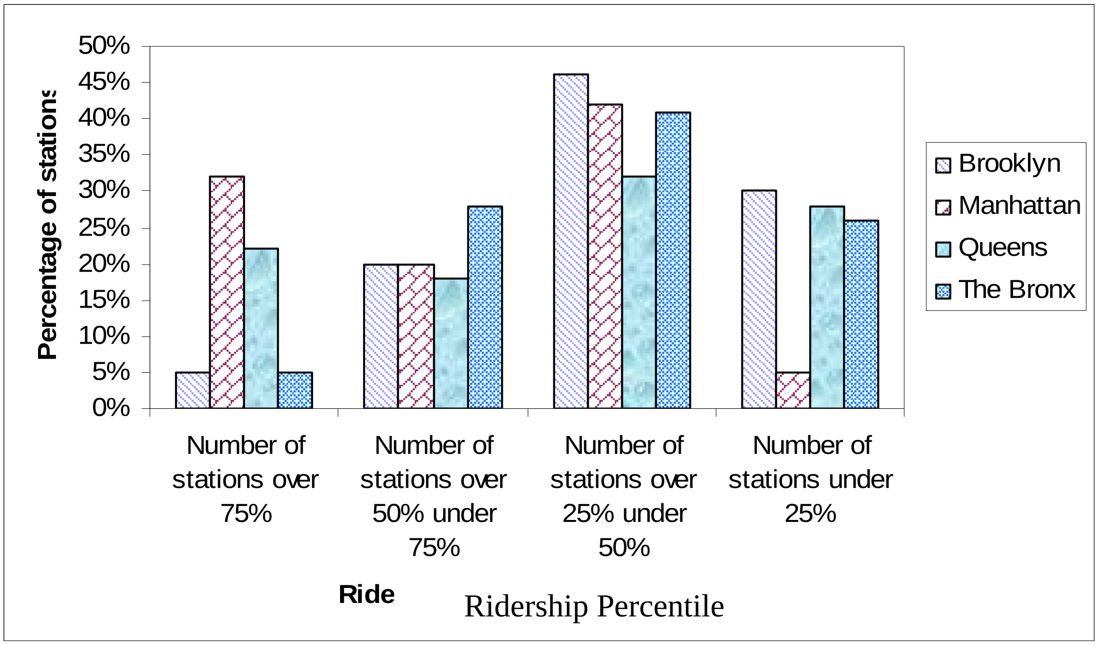
The Saturday ridership spatial distribution for four counties is shown in Figure A-7 in terms of absolute number of stations. Manhattan has the most number of stations which are above 75% percentile of station riderships in the city while Bronx has the fewest stations which are above 75% percentile. Brooklyn has the most number of stations which are under 25% percentile, while Manhattan has the fewest stations under 25% percentile. This pattern is consistent with that of weekday.

Figure A-7 Spatial Distribution of Subway Ridership in terms of the Number of Stations in each Borough that Falls in a Particular Percentile



The spatial distribution of ridership for four counties is shown in Figure A-8 in terms of percentage of stations in each borough. The stations in each percentile presented in Figure A-8 are calculated in Borough level. The overall pattern is almost the same as that of weekday. Manhattan has the highest percentage of stations which are above 75% percentile in ridership, while Brooklyn and Bronx both have lowest percentage (5%) of stations which are above 75% percentile in ridership. Only 5% of stations in Manhattan are under 25% percentile in ridership. Queens has 22% of stations which are above 75% percentile in ridership. In Saturday, stations in Manhattan still have the highest probability to have high ridership while stations in Brooklyn and Bronx have the lowest probability to have high ridership, which is exactly the same as that of weekday.

Figure A-8 Spatial Distribution of Station Ridership in terms of the Percentage of Stations in Each Borough that Falls in a Particular Percentile



A-3 Sunday Ridership

The histogram plot of Sunday total daily ridership is shown in Figure A-9. The X-axis and Y-axis have the same meanings as Figure A-1. On Sunday, most of stations' (about 95%) riderships fall under 12,000. The maximum daily ridership still occurs in Time square – 42nd Street station, which has 96,352 riders per day on average. On Sunday, the average ridership is less than that of Saturday. Other percentile information is shown in Table A-3. In Sunday, the mean total daily ridership is 5,476, with standard deviation to be 6,363. The 75% percentile of ridership is 6,323 and 25% percentile is 1,543. The ratios of each statistics in Sunday over that of weekday also fall into a narrow range, which varies from 0.41 to 0.48.

Figure 1 Sunday Daily Ridership Distribution

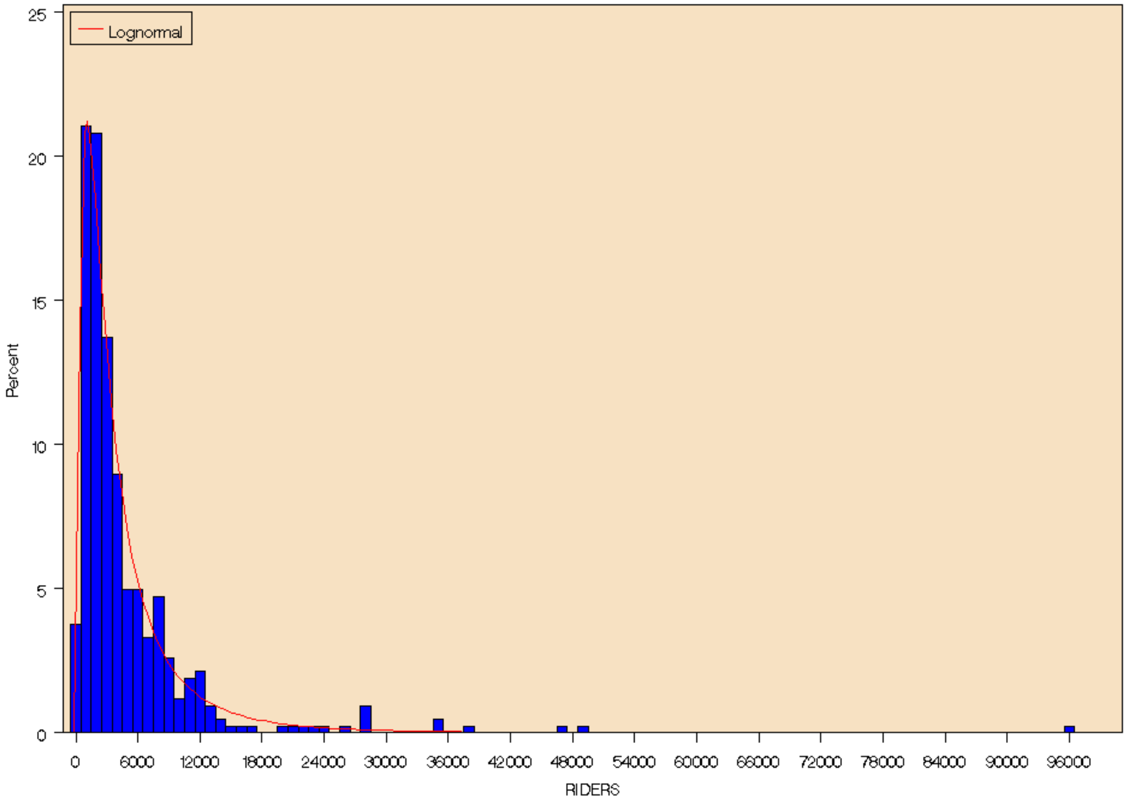
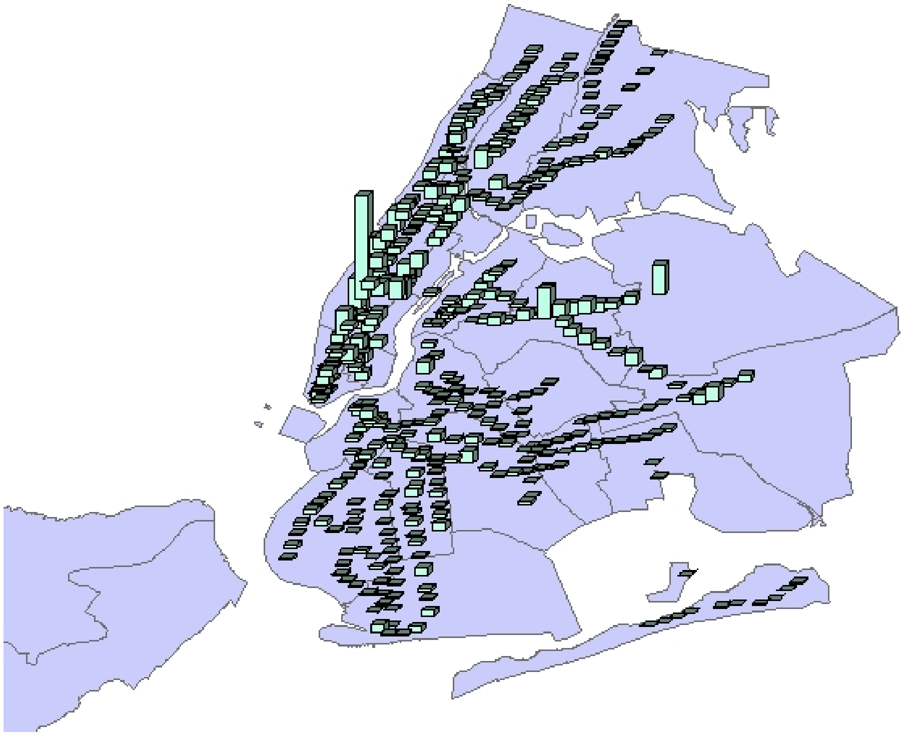


Table A-3 Statistical Description of Daily Ridership of New York City Subway

Statistic	Mean	Maximum	25 Percentile	50 Percentile	75 Percentile
Ridership	5,476	96,352	1,543	2,963	6,323

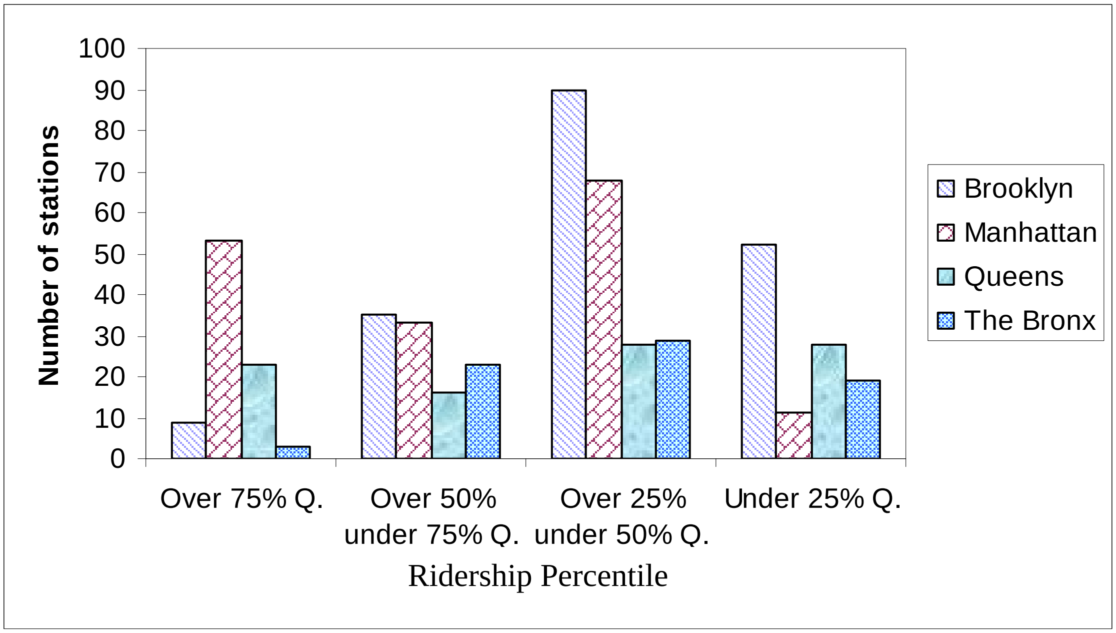
The spatial distribution of daily ridership for Sunday is shown in Figure A-10. Most of stations with higher ridership are located in Manhattan and Queens, while most of stations with lower ridership are located in Bronx and Brooklyn, which is the same as weekdays and Saturday.

Figure A-10 Overall Ridership Distribution For Sunday in May, 2005 by Stations



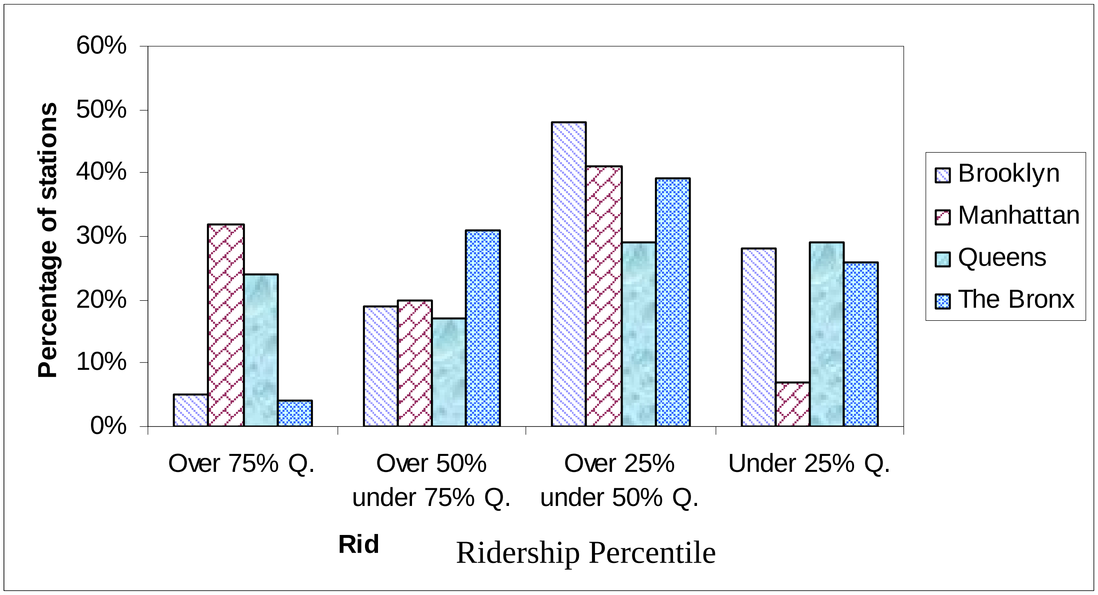
The spatial distribution of ridership for four counties in Sunday is shown in Figure A-11 in terms of absolute number of station. On Sunday, Manhattan still has the most number of stations which are above 75% percentile while Bronx has the fewest stations which are above 75% percentile. Brooklyn has the most number of stations which is under 25% percentile, while Manhattan has the fewest stations under 25% percentile. This pattern is observed both on weekdays and Saturday.

Figure A-11 Spatial Distribution of Subway Ridership in terms of the Number of Stations in each Borough that Falls in a Particular Percentile



The spatial distribution of ridership for four counties in Sunday is shown in Figure A-12 in terms of percentage of stations in each county. Manhattan still has the highest percentage (32%) of stations which are above 75% percentile, and about 7% of stations in Manhattan are under 25% percentile in ridership. Bronx has the lowest percentage (4%) of stations which are above 75% percentile in ridership. In all counties except for Queens, stations fall into 25% to 50% percentile occupy the highest percentage, while in Queens, stations fall into 0% to 25% percentile occupy the highest percentage.

Figure A-12 Spatial Distribution of Station Ridership in terms of the Percentage of Stations in Each Borough that Falls in a Particular Percentile



A-4 Summary

At a city level, most stations have a higher ridership on weekdays than those on weekends. Ridership decreases on Saturday and reaches the lowest level in Sunday. In general, we observe similarity in the spatial distribution of the daily ridership between weekdays and weekends. Manhattan has the most stations with high ridership while Bronx and Brooklyn have the fewest stations with high ridership. Queens comes in second after Manhattan in terms of the number of stations with high ridership. In particular, about 30% of Manhattan's stations have high ridership while only 5% of Bronx's and Brooklyn's stations have high ridership on weekdays and weekends. There are about 22% Queens' stations have high ridership on weekdays and weekends. On the other hand, Brooklyn, Bronx and Queens have quite similar percentage of stations with low ridership, about 30% while only 5% of Manhattan's stations have low ridership.