# Subway Ridership and Area Characteristics around Subway Stations in New York City

Hisayoshi Tanaka

Department of Geography, Hunter College, The City University of New York
and
New York City Department of City Planning, Transportation Division
Mailing address: 2 Lafayette Street, Suite 1200
New York, NY 10007
Phone: 212-442-4706
Fax: 212-442-4724
E-mail: htanaka@planning.nyc.gov

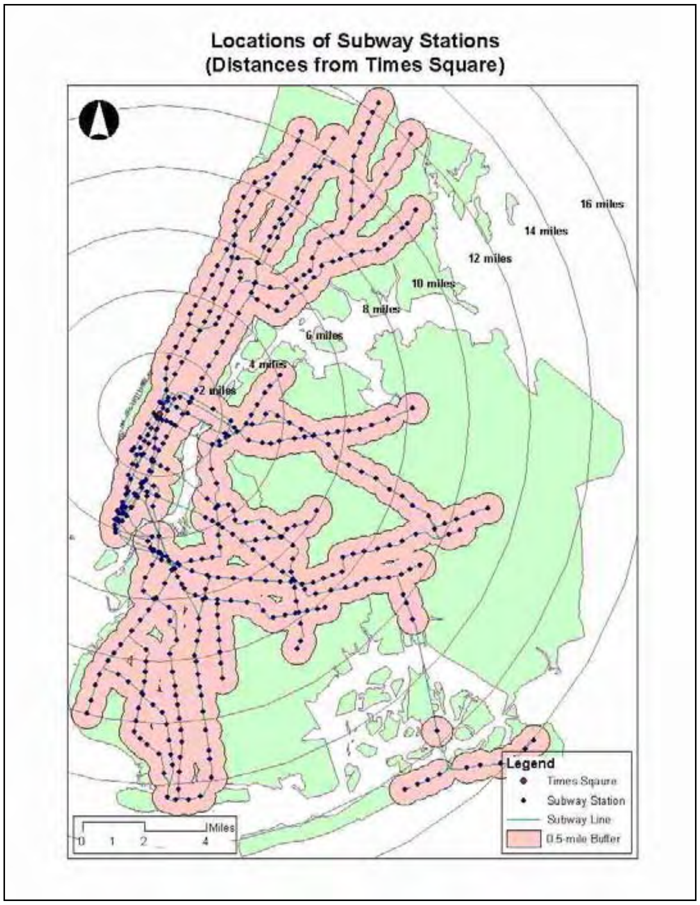
## Abstract:

This study is the preliminary summary of a currently ongoing project. The main purpose of the study is to examine how subway ridership and area characteristics around New York City (NYC) subway stations are correlated with each other. Average weekday subway ridership is used in analyses as a dependent variable with various independent variables from the following data categories: socio-economic characteristics, land use characteristics, and demographic characteristics. In addition, this study attempts to examine whether or not a subway station location can be a critical factor when such a correlation is estimated. As factors related to the location of subway stations, the following three variables are included in this study: 1) subway station locations by x- and y-coordinates, 2) distances from the center of the City (Times Square) to subway stations, and 3) distances from tax lots to their nearest subway stations. Analyses are performed with the following methodologies: 1) Ordinary Least Squares (OLS) linear regression and 2) Geographically Weighted Regression (GWR). OLS with the stepwise method identifies some influential variables to subway ridership. GWR is performed with these variables selected by OLS as well as a variable which represents subway station locations. The results of both methods are compared. As a result, GWR seems to reduce spatial non-stationarity.

## 1. Introduction:

Areas around subway stations (station areas) are often centers of all kinds of activities. For instance, people go to subway stations to use the subway/buses, or they go to station areas to do shopping. There is no doubt that the existence of subway station in a certain area provides people living nearby with more mobility and business opportunities. People also seem to consider the accessibility to public transportation as one of the most important factors when they look for a place to live. Thus, convenience will create additional intangible value to their lifestyles. Moreover, economic impacts caused by such convenience would not be only limited to that particular area. The impacts definitely go beyond a boundary of that area and even they will become vital to support a regional economic growth. Knowing the mechanism of how transportation would interact with non-transportation issues such as zoning, demographic distribution, etc. would be necessary to pursue in the continuous development of the entire region.

The primarily attempt of this study is to analyze how subway ridership and conditions of station areas are correlated each other in New York City (NYC) with using one of Geographically Weighted Regression (GWR) software packages.[1] Station areas for this study are defined areas which are within 0.50-mile buffers created around the subway stations, although the areas beyond this criterion could be still, somehow, influential to subway ridership (see Map1). However, a buffer size is determined, based on the concept of a “reasonable” walk distance. All relevant information to describe the station areas was gotten from various sources and a database was created to facilitate data manipulation.



Map 1: Locations of Subway Stations and Buffered Areas

There are over 400 subway stations in NYC. The dataset for this study consists of statistics such as socio-economic, demographic, and land use data from station areas of over 400 subway stations. Each subway station is indexed by a distance from Times Square (the center of the city) and tax lots around subway stations are also indexed, based on distances from the nearest subway stations.[2] In this way, the data would be analyzed in a two-dimensional way in terms of distances.

However, as the study went on, I realized that analyses on the entire NYC (excluding Staten Island) would have to deal with a huge dataset that easily exceeds the capacity of the GWR software package. Therefore, I aggregated the data to a reasonable size, based on the distances between tax lots and their nearest subway stations. A data aggregation process will be discussed in the following section, Data Aggregation.

## 2. Data and Software:

Data used for this study are of four kinds: transportation related data, socio-economic data, demographic data, and land use data. The details of each data are as follows:

Transportation related data:

- Average weekday subway ridership in year 2000.
- Subway station's geographic information (x, y-coordinates of subway station locations).[3]

Socio-economic data:

- Medium household income.

Demographic data:

- Total population.

Land use data (see Table 1):

- Lot area for each land use category (One & Two Family Buildings, Multi-Family Walkup Buildings, Multi-Family Elevator Buildings, Mixed Residential and Commercial

Buildings, Commercial and Office Buildings, Industrial and Manufacturing, Transportation and Utility, Public Facilities and Institutions, Open Space and Outdoor Recreation, Parking Facilities, Vacant Land, and Unknown)

Table 1: Types of Land Use

| Land Use Codes | Decodes |
|---|---|
| LU01 | One & Two Family Buildings |
| LU02 | Multi-Family Walk-up Building |
| LU03 | Multi-Family Elevator Building |
| LU04 | Mixed Residential and Commercial Buildings |
| LU05 | Commercial and Office Buildings |
| LU06 | Industrial and Manufacturing |
| LU07 | Transportation and Utility |
| LU08 | Public Facilities and Institutions |
| LU09 | Open Space and Outdoor Recreation |
| LU10 | Parking Facilities |
| LU11 | Vacant Land |
| LU12 | Unknown |

Note: Originally, no land use codes are assigned to tax lots whose categories are not identified. For a database creation purpose, “LU12: Unknown” is assigned to those tax lots.

Subway ridership data are gotten from the MTA New York City Transit (MTA NYCT) report, *Subway and Bus Ridership Report 2003*. US Census data are used for socio-economic and demographic information. As for land use information, the data are gotten from the Primary Land Use Tax Lot Output (PLUTO) data files, which are developed by the NYC Department of City Planning (NYCDCP).[4]

Based on the above available data, the following variables are created for further analyses.

- Average weekday subway ridership in year 2000 (AvgWkd00: persons)
- X-coordinate of subway station location (Long)
- Y-coordinate of subway station location (Latit)
- Distance from Time Square to each subway station (Dis_TS: miles)
- Distance from the nearest subway station to tax lot (Dis_ST: miles)
- Reciprocal of distance from the nearest subway station to tax lot (i_Dis_ST)
- Land use ratio (%) for each land use category (LU01, LU02, LU03, LU04, LU05, LU06, LU07, LU08, LU09, LU10, LU11, and LU12)
- Medium Household Income (MHHoldInc: dollars)
- Population Density (Den: persons per $Mi^2$)

Various software packages used for this study are SPSS and GWR3.0 for statistical work and ArcInfo for data organization and visualization.

## 3. Data Aggregation:

As shown in Table 2, there are approximately 400,000 tax lots in the 0.50-mile buffered areas around subway stations. Since the GWR software has the limitation on the number of observations that it can deal with, the following data aggregation process was taken to reduce the number of observations.

Table 2: Numbers of Tax Lots in Each Borough

| Borough | Number of Tax Lots |
|---|---|
| Manhattan | 42,886 |
| Bronx | 58,692 |
| Brooklyn | 209,391 |
| Queens | 103,968 |
| NYC (excluding Staten Island) | 414,937 |

Note: *Tax lots without x-y coordinates and sufficient information are excluded.

As the first step, tax lots which are located within the 0.5-mile buffered areas are selected from the entire PLUTO database. As long as the centroid of each tax lot is included in this buffered area, the entire tax lot, regardless of its size, is considered to belong to the buffered area. Then, distances between all tax lots (centroid of tax lot) and their nearest subway stations (x, y-coordinates of subway station location) are measured by ArcInfo. Based on the distances, tax lots are assigned to their nearest subway stations. Those allocated tax lots are also categorized into one of the following ring buffered areas, based on the distances:

- Ring 1: all tax lots within the area between 0.00 and 0.10 miles from each station.
- Ring 2: all tax lots within the area between 0.10 and 0.20 miles from each station.
- Ring 3: all tax lots within the area between 0.20 and 0.30 miles from each station.
- Ring 4: all tax lots within the area between 0.30 and 0.40 miles from each station.
- Ring 5: all tax lots within the area between 0.40 and 0.50 miles from each station.

Table 3 shows how many tax lots are included in each ring.

Table 3: Numbers of Tax Lots in NYC (Ring Buffer)

| Borough | Ring Buffer Size (Miles) | | | | | |
|---|---|---|---|---|---|---|
| | 0.0-0.1 | 0.1-0.2 | 0.2-0.3 | 0.3-0.4 | 0.4-0.5 | Total |
| Manhattan | 8,878 | 18,630 | 10,367 | 3,284 | 1,727 | 42,886 |
| Bronx | 5,167 | 15,993 | 16,230 | 11,850 | 9,452 | 58,692 |
| Brooklyn | 23,405 | 64,048 | 62,121 | 38,165 | 21,652 | 209,391 |
| Queens | 9,075 | 25,383 | 27,118 | 22,199 | 20,193 | 103,968 |
| Total | 46,525 | 124,054 | 115,836 | 75,498 | 53,024 | 414,937 |

Afterwards, attributes attached with tax lots assigned to each subway station are aggregated on a Ring-by-Ring basis. For instance, the 68th - Hunter College station on the Lexington line initially has 88, 267, 288, 113, and 79 tax lots for Rings 1, 2, 3, 4, and 5, respectively (see Table 4).

Table 4: Numbers of Tax Lots assigned to the 68th - Hunter College Station

| 68th - Hunter College Station | |
|---|---|
| Ring | Numbers of Tax Lots |
| 1 | 88 |
| 2 | 267 |
| 3 | 288 |
| 4 | 113 |
| 5 | 79 |
| Total | 835 |

Table 5: Numbers of Subway Stations assigned to Each Ring

| | |
|---|---|
| Ring | Numbers of Subway Stations |
| 1 | 468 |
| 2 | 466 |
| 3 | 435 |
| 4 | 363 |
| 5 | 296 |
| Total | 2,028 |

One representative figure for each Ring is created from tax lots selected for each Ring. Through this aggregation process, land use ratios are calculated - they express how large (in terms of percentage) each land use category is in each Ring. In the similar fashion, medium household income and population density for each Ring are also calculated.

As a result of the aggregation process, data from approximately 400,000 tax lots are shrunk into 2,028 observations (see Table 5). They are used for analyses of this study.

## 4. Assumptions:

Before analyses, the following assumptions are made:

- Subway ridership at a particular subway station is influenced by the only attributes whose tax lot is assigned to that subway station - there are no tax lots which can be influential to two or more subway stations.
- To measure distances between subway stations and tax lots, Euclidean distances are applied to two geocoded points: the location of subway station and the centroid of tax lot. Regardless of the size of tax lots, distances are simply dependent on those two points.
- To simplify, tax lots included in a particular Ring are considered to have the same distance from their nearest subway station. The distance is the simple average of all distances assigned to tax lots in the Ring under consideration.
- Population and medium household income are aggregated on a Ring-to-Ring basis and these statistics are assumed to be uniformly distributed in each Ring.

## 5. Results:

As the first attempt, ordinary regression analysis with the stepwise method was performed to reduce the number of variables. This task was done by SPSS. As the second step, the GWR software package was used to see how a geographical factor (location of subway stations) would change the results by the ordinary regression. Only those variables selected by the process of the stepwise method were included in this analysis.

Basic descriptive statistics for all used variables are summarized in Tables 6-1 and 6-2.

Table 6-1: Descriptive Statistics for Land Use Variables

| Land Use Code | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| LU01 | 2,028 | .0000 | 100.0000 | 20.707094 | 22.8560242 |
| LU02 | 2,028 | .0000 | 100.0000 | 14.671567 | 13.9523973 |
| LU03 | 2,028 | .0000 | 100.0000 | 11.416721 | 16.6685784 |
| LU04 | 2,028 | .0000 | 100.0000 | 8.652294 | 11.2299185 |
| LU05 | 2,028 | .0000 | 100.0000 | 10.662634 | 18.1543552 |
| LU06 | 2,028 | .0000 | 100.0000 | 6.588207 | 14.0729379 |
| LU07 | 2,028 | .0000 | 100.0000 | 3.893978 | 11.0233911 |
| LU08 | 2,028 | .0000 | 100.0000 | 9.444404 | 13.6907174 |
| LU09 | 2,028 | .0000 | 100.0000 | 5.661601 | 16.2507951 |
| LU10 | 2,028 | .0000 | 69.7395 | 2.898807 | 5.0893895 |
| LU11 | 2,028 | .0000 | 100.0000 | 4.553136 | 10.2716258 |
| LU12 | 2,028 | .0000 | 100.0000 | .800248 | 4.5188265 |
| Valid N (listwise) | 2,028 | | | | |

Table 6-2: Descriptive Statistics for Non-Land Use Variables

| | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| AvgWkd00 | 2,028 | 78 | 153,505 | 12,458.79 | 20,488.638 |
| Den | 2,028 | .0000 | 217,136.4195 | 52,025.367068 | 33,661.7807828 |
| Dis_TS | 2,028 | .0792 | 16.0631 | 6.343591 | 3.3417342 |
| Dis_ST | 2,028 | .0000 | .4639 | .215809 | .1268741 |
| I_Dis_ST | 2,027 | 2.1558 | 63.1900 | 7.572163 | 6.1920386 |
| MHHoldInc | 2,028 | .00 | 143,438.67 | 38,987.2896 | 19,879.91491 |
| Valid N (listwise) | 2,027 | | | | |

Average weekday subway ridership in year 2000 (AvgWkd00) was used as a dependent variable, while other variables were included in a regression equation as independent variables. The equation for the ordinary regression is as follows:

$$AvgWkd00 = \alpha_0 + \sum \alpha_i{*}LU_i + \alpha_{13}{*}Den + \alpha_{14}{*}Dis\_TS + \alpha_{15}{*}i\_Dis\_ST + \alpha_{16}{*}MHHoldInc$$

where i = 1 to 12.

The results of the ordinary regression analysis with the stepwise method are summarized in Tables 7 and 8.

The stepwise method selects seven variables among many as significantly influential variables to subway ridership. The results show that subway ridership is positively correlated to Commercial and Office Buildings land use (LU05), medium household income (MHHoldInc), and population density (Den), while it is negatively correlated to Industrial and Manufacturing and Multi-Family Walk-up Buildings land uses (LU06 and LU02, respectively), distance from the center (Dis_TS), and the reciprocal of distance from subway station (i_Dis_ST) (see Table 8). The details are as follows:

- One-percent increase in a size of the Commercial and Office Buildings land use (LU05) may result in 422-person increase in subway ridership.
- One-percent increase in a size of the Industrial and Manufacturing land use (LU06) may result in 110-person decrease in subway ridership.

- One-percent increase in a size of the Multi-Family Walk-up Buildings land use (LU02) may result in 144-person increase in subway ridership.
- 100-dollar increase in medium household income (MHHoldInc) may result in 101-person increase in subway ridership.
- 100-person increase in density (Den) may result in 27-person increase in subway ridership.
- One-mile increase in distance from Times Square (Dis_TS) may result in 2,034-person decrease in subway ridership.
- One increase in the reciprocal of distance from each subway station (i_Dis_ST) may result in 316-person decrease in subway ridership. This can be interpreted as follows: a tax lot located closer to a subway station may have some potential to generate more subway ridership.

Table 7: Results of Ordinary Regression Analysis with Stepwise Method (Model Summary)

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .498 | .248 | .247 | 17,779.938 |
| 2 | .600 | .360 | .359 | 16,404.362 |
| 3 | .610 | .372 | .371 | 16,257.718 |
| 4 | .615 | .378 | .377 | 16,172.360 |
| 5 | .619 | .383 | .381 | 16,117.369 |
| 6 | .624 | .390 | .388 | 16,031.575 |
| 7 | .625 | .391 | .389 | 16,018.003 |

Table 8: Results of Stepwise Method (Coefficients)

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 6,464.011 | 458.060 | | 14.112 | .000 |
| | LU05 | 561.654 | 21.755 | .498 | 25.817 | .000 |
| 2 | (Constant) | 21,468.495 | 901.699 | | 23.809 | .000 |
| | LU05 | 439.344 | 21.096 | .389 | 20.826 | .000 |
| | Dis_TS | -2,159.117 | 114.618 | -.352 | -18.837 | .000 |
| 3 | (Constant) | 15,502.867 | 1,320.284 | | 11.742 | .000 |
| | LU05 | 406.440 | 21.584 | .360 | 18.831 | .000 |
| | Dis_TS | -1,929.822 | 119.578 | -.315 | -16.139 | .000 |
| | MHHoldInc | .125 | .020 | .121 | 6.138 | .000 |
| 4 | (Constant) | 17,352.674 | 1,370.247 | | 12.664 | .000 |
| | LU05 | 447.867 | 23.185 | .397 | 19.317 | .000 |
| | Dis_TS | -1,925.184 | 118.955 | -.314 | -16.184 | .000 |
| | MHHoldInc | .123 | .020 | .120 | 6.102 | .000 |
| | i_Dis_ST | -299.505 | 63.266 | -.090 | -4.734 | .000 |
| 5 | (Constant) | 19,177.481 | 1,445.510 | | 13.267 | .000 |
| | LU05 | 440.567 | 23.184 | .390 | 19.003 | .000 |
| | Dis_TS | -2,044.580 | 122.539 | -.333 | -16.685 | .000 |
| | MHHoldInc | .116 | .020 | .112 | 5.733 | .000 |
| | i_Dis_ST | -304.072 | 63.062 | -.092 | -4.822 | .000 |
| | LU06 | -101.385 | 26.335 | -.070 | -3.850 | .000 |
| 6 | (Constant) | 23,006.494 | 1,647.272 | | 13.966 | .000 |
| | LU05 | 411.204 | 23.870 | .364 | 17.227 | .000 |
| | Dis_TS | -2,149.844 | 123.874 | -.351 | -17.355 | .000 |
| | MHHoldInc | .098 | .020 | .095 | 4.783 | .000 |
| | i_Dis_ST | -300.419 | 62.731 | -.091 | -4.789 | .000 |
| | LU06 | -134.828 | 27.119 | -.093 | -4.972 | .000 |
| | LU02 | -132.948 | 27.911 | -.091 | -4.763 | .000 |
| 7 | (Constant) | 20,767.691 | 1,960.044 | | 10.596 | .000 |
| | LU05 | 421.887 | 24.385 | .374 | 17.301 | .000 |
| | Dis_TS | -2,033.680 | 135.532 | -.332 | -15.005 | .000 |
| | MHHoldInc | .101 | .021 | .098 | 4.917 | .000 |
| | i_Dis_ST | -316.481 | 63.142 | -.096 | -5.012 | .000 |
| | LU06 | -110.211 | 29.516 | -.076 | -3.734 | .000 |
| | LU02 | -143.921 | 28.371 | -.098 | -5.073 | .000 |
| | Den | .027 | .013 | .044 | 2.103 | .036 |

a Dependent Variable: AvgWkd00

These selected variables were used for the GWR analysis. The equation used for this task was as follows:

$$\text{AvgWkd00} = \alpha_0 + \alpha_2 * LU_2\,(g) + \alpha_5 * LU_5\,(g) + \alpha_6 * LU_6\,(g) + \alpha_{13} * \text{Den}\,(g) + \alpha_{14} * \text{Dis\_TS}\,(g) + \alpha_{15} * \text{i\_Dis\_ST}\,(g) + \alpha_{16} * \text{MHHoldInc}\,(g)$$

where (g) is a location factor.

The results of the GWR analysis are available in Appendix A. Table 9 shows the results of the global regression model extracted from Appendix A. These results are very much similar to those by SPSS (see Tables 8 and 9).

Table 9: Results of Global Model and Median Values for Each Parameter by GWR

| | Global Model | | | GWR | | |
|---|---|---|---|---|---|---|
| | Estimate | Std Err | T | Minimum | Median | Maximum |
| (Constant) | 20,754.495 | 1,958.733 | 10.596 | -81,333.694 | 3,680.829 | 171,899.353 |
| LU02 | -143.799 | 28.359 | -5.071 | -2,749.263 | -27.756 | 597.085 |
| LU05 | 422.081 | 24.364 | 17.324 | -106.641 | 81.302 | 778.996 |
| LU06 | -109.990 | 29.493 | -3.729 | -2,606.386 | -15.219 | 2,664.846 |
| Den | .027 | .013 | 2.114 | -.360 | .032 | .507 |
| i_Dis_ST | -316.220 | 63.117 | -5.010 | -1,905.659 | -64.996 | 771.061 |
| Dis_TS | -2,032.777 | 135.442 | -15.008 | -99,834.947 | -38.183 | 32,770.283 |
| MHHoldInc | .101 | .020 | 4.913 | -1.506 | .025 | 1.334 |

However, if a location factor is included in the analyses, the results are improved. For instance, the value of Akaike Information Criterion (AIC) was 45,032 for the global model, while it reduced to 43,639 for the GWR model. In the same way, adjusted $R^2$ is also improved dramatically from 0.39 (global model) to 0.74 (GWR model) (see Table 10). The results of ANOVA show that F-test rejected the null hypothesis, $H_0$: OLS and GWR are the same (see Table 11).

Table 10: Summary of Test Statistics

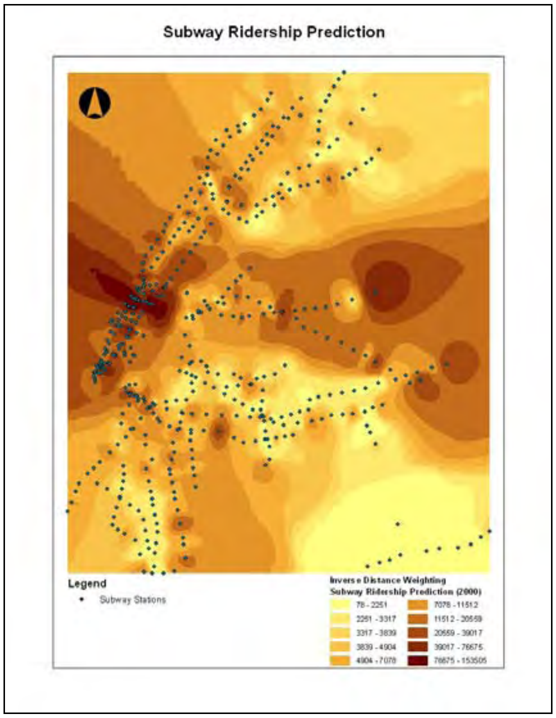
| | Global Regression | GWR (Estimation) |
|---|---|---|
| Residual sum of squares | 518,041,208,578.980220 | 184,499,406,907.040100 |
| Effective number of parameters | 8.000000 | 304.067787 |
| Sigma | 16,014.245029 | 10,345.164356 |
| Akaike Information Criterion | 45,032.372844 | 43,639.144286 |
| Coefficient of Determination | .391186 | .783172 |
| Adjusted R-square | .388774 | .744906 |

Table 11: ANOVA

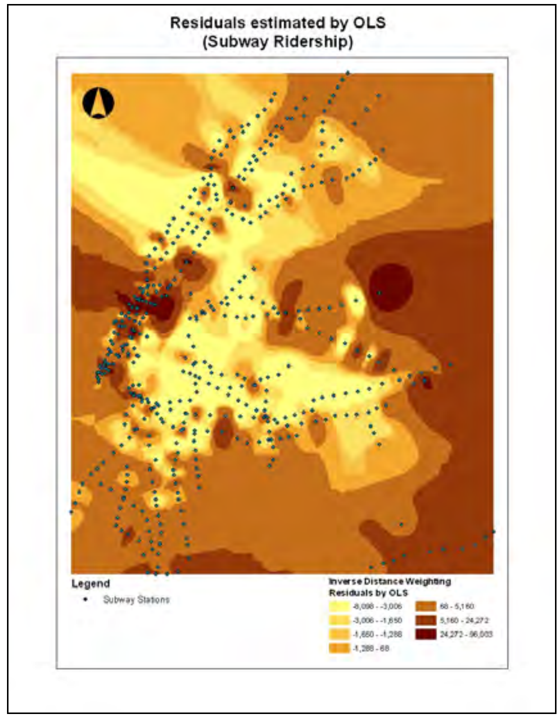
| Source | SS | DF | MS | F |
|---|---|---|---|---|
| OLS Residuals | 518,041,208,579.0 | 8.00 | | |
| GWR Improvement | 333,541,801,984.0 | 296.07 | 1,126,572,415.6467 | |
| GWR Residuals | 184,499,406,907.0 | 1723.93 | 107,022,425.5424 | 10.5265 |

Although test statistics were improved by GWR, the results of GWR may not be easily interpreted because estimated values for parameters are given in ranges, which are defined by minimum, median, and maximum (see Table 9).

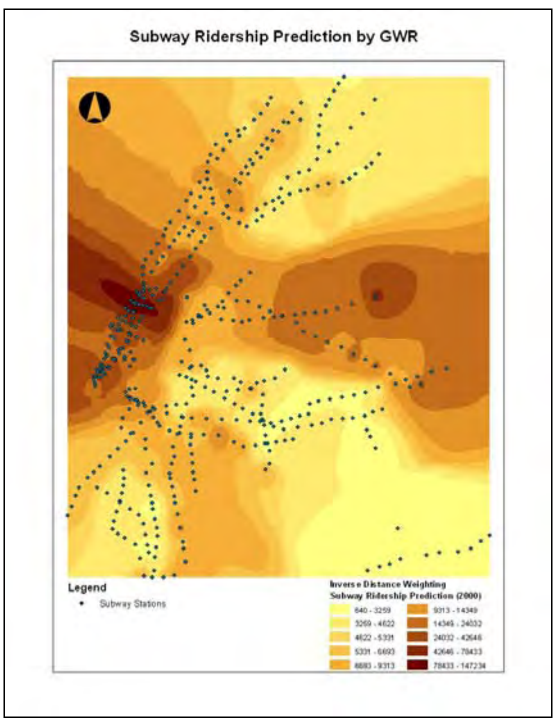
Subway ridership predicted by GWR (used Inverse Distance Weighting with the optimized power value to visualize (IDW Opt), Map 4) seems to be more smoothed out, compared with one predicted with the original subway ridership by IDW Opt (Map 2), although both maps seem to have similar patterns. Since both maps do not have the same scales of contours, it cannot be easily compared. However, those maps, at least, would imply that higher subway ridership is expected in areas such as Midtown (Manhattan) and Flushing (Queens). Map 2 shows that there are also a couple of areas in the Bronx and Brooklyn, where a large volume of subway ridership is expected. The original subway ridership data seem to be noisy with spatial non-stationarity. In fact, based on residuals estimated by OLS, a residual map is created by IDW Opt to observe the existence of spatial non-stationarity (Map 3).



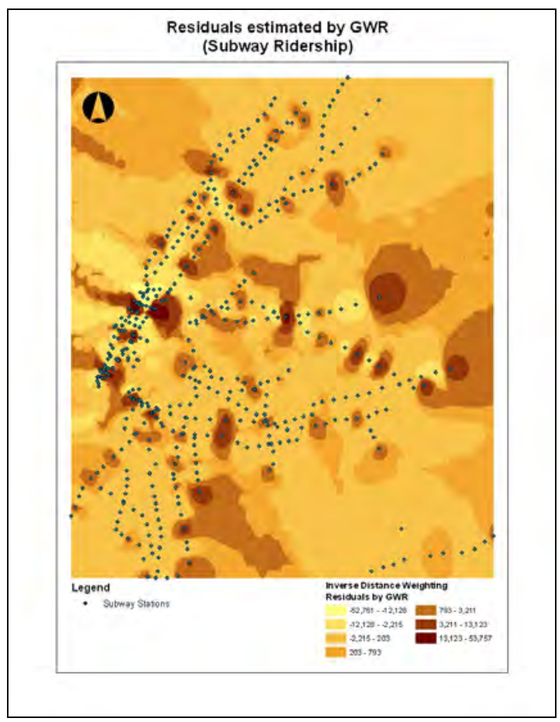

Map 2: Predicted Subway Ridership by OLS




Map 3: Residuals estimated by OLS




Map 4: Predicted Subway Ridership by GWR




Map 5: Residuals estimated by GWR

After area characteristics are included in the analyses, those areas with higher subway ridership than other station areas seem to be merged into their neighborhoods and almost phased out (Map 4). The residual map for the GWR model is also created (Map 5). Map 5 shows that GWR seems to reduce spatial non-stationarity because areas with larger residuals are more likely distributed in the entire NYC randomly.

Due to differences in scales of contours on Maps 3 and 5, it is not easy to compare residuals on both maps. However, both maps would be sufficient to show which case is more contaminated with spatial non-stationarity.

## 6. Conclusions:

The purpose of this study is to examine how subway ridership and characteristics of station areas are correlated with each other. Average weekday subway ridership is used in the analyses as a dependent variable with various independent variables from the following data categories: socio-economic characteristics, land use characteristics, and demographic characteristics.

The OLS regression analysis with the stepwise method was initially performed to eliminate insignificant variables from a full model regression equation. It was able to identify some variables which would be more influential to subway ridership. These identified variables were used for the GWR analysis. They are as follows: Multi-Family Walk-up Buildings land use (LU02), Commercial and Office Buildings land use (LU05), Industrial and Manufacturing land use (LU06), medium household income (MHHoldInc), population density (Den), distance from the center (Dis_TS), and distance from subway station (i_Dis_ST).

The results point out that land use characteristics around station areas (especially, how much station areas are designated to Commercial and Office Buildings land use) seem to play more important roles to estimate subway ridership. If station areas are highly designated to Commercial and Office Buildings land use, subway stations in such station areas seem to be able to generate more subway ridership. Even compared with other variables, Commercial and Office Buildings land use seems to be the variable which can influence subway ridership more effectively. Higher medium household income also seems to result in more subway ridership. However, since the level of income is solely used as a socio-economic condition in this study, this result might not be sufficient to support that the richer use more public transportation. Rather, the result may imply that wealthy neighborhoods (areas dominated by people with higher income) may be more attractive to many. Therefore, even people who do not live nearby may come over to such neighborhoods. As a result, subway stations located nearby wealthy neighborhoods have higher subway ridership. There might be some room to think that the subway may be used by all kinds of people, regardless of their social statuses (the subway could be the real *public* transportation to all the New Yorkers). However, this study did not include any variables to examine such a hypothesis.

In this study, data are collected from areas buffered with the 0.5-mile radius around each subway station. The buffered areas were further broken down into five Rings. However, there is some room for a discussion whether the 0.5-mile radius for station areas and the 0.1-mile width for Rings were enough sizes to catch all activities around subway stations and to display trends in station areas. Data from much wider buffered areas or narrower ring buffered areas may lead to different results. To solve this problem, a further analysis will be required.

At last, as seen in Map 3, it is obvious that the original subway ridership data themselves are contaminated with spatial non-stationaity. Huge spikes of residuals are observed in areas such as Midtown Manhattan, Lower Manhattan, and Flushing (Queens). However, GWR with area characteristic variables successfully reduces spatial non-stationarity (Map 5).

[1] The software used for Geographically Weighted Regression is called GWR3.0, which is developed by Martin Charlton, Stewart Fotheringham, and Chris Brunsdon. A user manual is available at the following web address: http://www.ncrm.ac.uk/publications/methodsreview/MethodsReviewPaperNCRM-006.pdf

[2] It is assumed in this study that Times Square is the center of the city. Reasons why Times Square is chosen are that Times Square is accessible by the most subway lines and that it is actually the center of business activities. Distances between subway stations and this alleged center are measured and used for the further analyses.

[3] GIS shapefiles for subway stations and lines, created in 2002, are used to obtain necessary geographic data for this study.

[4] The Primary Land Use Tax Lot Output (PLUTO) data files contain over 70 data fields derived from extracts of mainframe data files maintained by NYC Department of City Planning (NYCDCP) and other agencies. The PLUTO data files consist of the following four basic types of data: 1. tax lot characteristics, 2. building characteristics, 3. geographic/political/administrative districts, and 4. geographic fields formatted for use with NYCDCP geographic mapping files.

# Appendix A
# Results of GWR Analysis

```
** Dependent (y) variable..........AvgWkd00
** Easting  (x-coord) variable.....LONG
** Northing (y-coord) variable.....LATIT
** No weight variable specified
** Independent variables in your model...
 LU02     LU05     LU06     Den   i_Dis_ST Dis_TS  MHHoldInc
** Kernel type:  Adaptive
** Kernel shape: Bi-Square
** Bandwidth selection by AICc minimisation
** Use all regression points
** Calibration history requested
** Prediction report requested
** Output estimates to be written to .e00 file
** No significance test for spatial variation
** Casewise diagnostics to be printed

*** Analysis method ***
*** Geographically weighted multiple regression
** Cartesian coordinates: Euclidean Distance
***************************************************************
*                                                             *
*        GEOGRAPHICALLY WEIGHTED GAUSSIAN REGRESSION          *
*                                                             *
***************************************************************
 Number of data cases read: 2028
 Observation points read...

 Dependent mean=  12458.7539
 Number of observations, nobs= 2028
 Number of predictors,   nvar= 7
 Observation Easting extent:   274955.
 Observation Northing extent:  326690.
*Finding bandwidth...
  ... using all regression points
 This can take some time...
 *Calibration will be based on  2028 cases
*Adaptive kernel sample size limits:     101   2028
 *AICc minimisation begins...
           Bandwidth                    AICc
        696.475749365000       44482.788057697122
       1064.500000000000       44631.008373585377
        469.024253328291       44380.384917708659
```

```
        328.451497701254       44277.098983414173
        241.572756655782       44177.393814528703
        187.878741681272       44091.034844997557
        154.694015367456       43977.384144314114
        134.184726556670       43872.130409869183
        121.509288960878       43756.023636277117
        113.675437688554       43673.485368321548
        108.833851329654       43628.787315235670
        105.841586394333       43606.036396112904
        103.992264957219       43568.945507202399
** Convergence after    13 function calls
** Convergence: Local Sample Size=    104

**********************************************************
*           GLOBAL REGRESSION PARAMETERS                 *
**********************************************************
Diagnostic information...
Residual sum of squares......... 518041208578.980220
Effective number of parameters..           8.000000
Sigma...........................       16014.245029
Akaike Information Criterion....       45032.372844
Coefficient of Determination....           0.391186
Adjusted r-square...............           0.388774

Parameter              Estimate               Std Err                      T
---------          ------------          ------------           ------------
Intercept     20754.494953946290      1958.732924633875       10.595877647400
 LU02           -143.799111228895        28.359445415887        -5.070589542389
 LU05            422.081497367482        24.364280370032        17.323781967163
 LU06           -109.990288040628        29.493190877368        -3.729345083237
 Den               0.026924642148         0.012736748416         2.113933801651
 i_Dis_ST       -316.220339622047        63.116621445867        -5.010096073151
 Dis_TS        -2032.777268683483       135.442173689481       -15.008451461792
 MHHoldInc         0.100694997855         0.020496914902         4.912690639496


**********************************************************
*                GWR ESTIMATION                          *
**********************************************************
 Fitting Geographically Weighted Regression Model...
 Number of observations............ 2028
 Number of independent variables... 8
  (Intercept is variable 1)
 Number of nearest neighbours...... 104
 Number of locations to fit model.. 2028

Diagnostic information...
```

```
Residual sum of squares......... 184499406907.040100
Effective number of parameters..         304.067787
Sigma...........................       10345.164356
Akaike Information Criterion....       43639.144286
Coefficient of Determination....           0.783172
Adjusted r-square...............           0.744906
```

```
**********************************************************
*                      ANOVA                             *
**********************************************************
```

| Source | SS | DF | MS | F |
|---|---|---|---|---|
| OLS Residuals | 518041208579.0 | 8.00 | | |
| GWR Improvement | 333541801984.0 | 296.07 | 1126572415.6467 | |
| GWR Residuals | 184499406907.0 | 1723.93 | 107022425.5424 | 10.5265 |

```
**********************************************************
*          PARAMETER 5-NUMBER SUMMARIES                  *
**********************************************************
```

| Label | Minimum | Lwr Quartile | Median | Upr Quartile | Maximum |
|---|---|---|---|---|---|
| Intrcept | -81333.693907 | -3019.579495 | 3680.828718 | 14035.759554 | 171899.353327 |
| LU02 | -2749.263137 | -74.553194 | -27.756116 | 6.845055 | 597.084750 |
| LU05 | -106.641006 | 32.718527 | 81.301696 | 202.110722 | 778.996344 |
| LU06 | -2606.386092 | -63.943464 | -15.219386 | 23.165072 | 2664.846337 |
| Den | -0.360307 | 0.005514 | 0.032245 | 0.085798 | 0.507218 |
| i_Dis_ST | -1905.658593 | -211.833740 | -64.995591 | -14.689082 | 771.060504 |
| Dis_TS | -99834.947116 | -1574.989067 | -38.183341 | 1041.620627 | 32770.282954 |
| MHHoldInc | -1.506462 | -0.043657 | 0.024525 | 0.127357 | 1.334036 |

| | <----------------- LOWER | ----------------- | ----------------> | <----------------- | UPPER ----------------- | | ----------------> |
|---|---|---|---|---|---|---|---|
| Label | Far Out | Outer Fence | Outside | Inner Fence | Inner Fence | Outside | Outer Fence | Far Out |
| Intrcept | 29 | -54185.596642 | 63 | -28602.588069 | 39618.768127 | 84 | 65201.776701 | 111 |
| LU02 | 185 | -318.747940 | 67 | -196.650567 | 128.942428 | 34 | 251.039801 | 41 |
| LU05 | 0 | -475.458057 | 0 | -221.369765 | 456.199014 | 69 | 710.287306 | 7 |
| LU06 | 126 | -325.269074 | 102 | -194.606269 | 153.827877 | 93 | 284.490682 | 120 |
| Den | 43 | -0.235337 | 75 | -0.114911 | 0.206223 | 52 | 0.326648 | 80 |
| i_Dis_ST | 61 | -803.267714 | 116 | -507.550727 | 281.027905 | 46 | 576.744893 | 35 |
| Dis_TS | 152 | -9424.818149 | 44 | -5499.903608 | 4966.535168 | 69 | 8891.449709 | 73 |
| MHHoldInc | 79 | -0.556700 | 42 | -0.300179 | 0.383879 | 76 | 0.640400 | 36 |