SPATIO-TEMPORAL PATTERN ANALYSIS OF TAXI TRIPS IN NEW YORK CITY

Hartwig H. Hochmair
University of Florida, Geomatics Program
3205 College Avenue
Ft. Lauderdale, FL-33314
Tel: 954 577-6317 Fax: 954 475-4125; E-mail: hhochmair@ufl.edu

Word Count: 4977 words text + 8 tables/figures x 250 words (each) = 6977 words

Revised version submitted: March 11, 2016
**ABSTRACT**

A growing number of extensive datasets provide transportation planners with the necessary means to analyze urban travel patterns and gain insight into urban dynamics. This paper explores the spatial and temporal variation of taxi trips in New York City (NYC) by analyzing 29 million trip records from a freely available dataset. The study examines the role of airports in trip generation and attraction, as well as the variation of travel speed during the day. Comparison of hourly trip frequencies between weekday and weekend in each district reveals similarities and differences in functional drivers of taxi trip demand. A negative binomial regression model is presented which predicts the number of taxi trips per district from subway, train, and bus infrastructure, as well as socio-economic and land use variables, where eigenvector spatial filtering is applied to explicitly model spatial autocorrelation. Independent of predictor variables, a combination of subway ridership and taxi trip numbers for each district in a modemix variable allows, using Local Indicators of Spatial Association statistics, to identify districts that exhibit an increased inclination towards taxi use and are currently poorly served by public transit. This approach could be used as a decision support tool for deciding where investments in rapid transit infrastructure and service would be particularly effective in order to increase transit mode share.

*Keywords: travel demand, taxi trips, subway ridership, open data, GIS*
INTRODUCTION

Taxi service can be viewed as a higher level of public transit by providing personal, on demand, point-to-point transportation. It is used by travelers whose value of time and requirements on comfort are higher than those of bus, metro, and train uses, or who are in need of such a service, such as (a) elderly, young or low income people without their own vehicle who live in areas that are poorly served by public transit, (b) people with disabilities needing door-to-door service, (c) tourists and visitors who choose not to drive in an unfamiliar city or to use its public transit system (1), or (d) people not owning a car undertaking a trip which cannot be easily done by transit, e.g. when buying heavy grocery. This study explores spatial and temporal variations in taxi trip demand in New York City based on a freely available set of taxi trip data, and extends methods and results from related studies that were using similar taxi ridership datasets. A novel aspect of this paper is the application of a spatial negative binomial regression model using eigenvector spatial filtering (ESF) to explore the relationship between taxi trips and explanatory variables, such a transit infrastructure, while explicitly considering spatial autocorrelation in observations. Another novel aspect is the application of local statistical techniques in a combined variable of taxi trip counts and subway boarding numbers, which can help to identify areas that are currently poorly served by public transit.

Data used in the study come from the Taxi and Limousine Commission of New York City (TLC) and contain information about all medallion taxi trips in NYC in 2013. This freely available data is an example of a growing number extensive datasets that provide transportation planners with the opportunity to better understand urban travel patterns and urban dynamics. Open data is based on the idea that certain data should be freely available to everyone to use and republish, without restrictions from copyright or patents (2). NYC provides currently access to over 1350 data sets on the Web (https://data.cityofnewyork.us/dashboard) in a variety of machine-readable formats. The New York City Council requires all city agencies to open their data by 2018.

PREVIOUS WORK

Taxicab is a significant transportation mode in urban areas since it complements other public transport modes with flexible door-to-door service through uninterrupted service (3, 4). Therefore, a thorough understanding of taxi trip behavior and its relationship to transit infrastructure is important to further increase the efficiency of urban transportation systems. Several recent studies focus on improving taxi efficiency. A study conducted in Taipei City showed that 60–73% of their operation hours, taxi drivers were driving without passengers because they did not know where potential customers were (5). This problem was addressed in the paper by mining historical data to predict passenger demand distributions with respect to time, weather, and taxi location, using clustering algorithms. Another study applied time series forecasting techniques on data from taxis equipped with real-time vehicle location systems to make short term predictions on the passenger demand for 63 taxi stands in the city of Porto, Portugal (6). A predictive model for the number of vacant taxis in a given area based on time of the day, day of the week, and weather condition, using 2.6 million anonymous locations of 150 taxis in Lisbon, Portugal, is presented in (7).

A wide range of spatial data sources have recently emerged due to advances in sensor and telecommunication technology, informatics, and data processing, which facilitates the analysis of various aspects of travel behavior. These include among others GPS tracking systems, which are used for various transportation modes including cycling, walking, bus, and taxi. GPS tracking is also used in all New York City taxis, which are regulated by the TLC. Over the past few years, the TLC provided access to its trip database which contains millions of taxi trips with temporal and
spatial information, including taxi pickup and drop-off date, time, and location, as well as fare and distance traveled.

Recent research studies used this data source to analyze various aspects of taxi ridership in NYC. One study found that travel times from truck-GPS data can be better estimated from taxi-GPS data during AM and PM periods than during night time, which indicates that speed differences between taxis and trucks are greater for free-flow conditions (8). Another study used 10 months of 2010 NYC taxi trip data to estimate a multiple linear regression model for each hour of the day to model pickups and drop-offs (9). The results identified six important explanatory variables for taxi trips, which include population, education, age, income, transit access time, and employment. A related study extracted one-week NYC taxi trip data from 2009 and predicted trip frequency through six variables in Ordinary Least Square (OLS) and geographically weighted regression (GWR) models (4). Identified significant variables include commuting time, education, income, proportion of commercial area, road density, and subway accessibility. The GWR model demonstrated a strong spatial variability for parameter estimations and outperformed the OLS model in terms of model fit and mitigation of spatial autocorrelation in residuals. The effect of winter storms on taxi trip characteristics was explored in another study, showing that travelers opted to take shorter trips during the snowstorm than during fair weather conditions (10).

Another paper discusses the anticipated impacts of two recently enacted taxi regulation changes on revenue increase, which are a fare increase, and the use of smart phone taxi applications (11). The study concludes that the fare increase is not expected to increase driver incomes equally for all periods, and that the apps’ effectiveness depends on the time of day and weather conditions. Using a classification and regression tree model on average trip speed per mile obtained from observed taxi trips it was shown that travel time reliability-related Day-of-Week and Time-of-Day periods did not follow the conventional understanding of morning and evening, peak, and midday travel time patterns (12). Using the notion of a shareability network, observed taxi trips were used to model the collective benefits of taxi sharing as a function of passenger inconvenience, and to efficiently compute optimal sharing strategies on massive datasets (13). This simulation study has, however, not yet been tested with real passengers, hence it is not yet clear what constitutes the maximum acceptable passenger discomfort imposed by sharing a cab and caused delays.

Other analysis tasks that utilized NYC taxi trip data provided by the TLC include a binary logit model to predict the mode choice between transit and taxi mode (14), comparison of trip characteristics between summer (July) and non-summer (March) months (15), and the development of TaxiVis, which is an analysis environment that allows users to visually query taxi trips under consideration of spatial, temporal, and attribute constraints (16).

STUDY SETUP

Study area and data sources
The study area is New York city which covers five counties (Bronx, Kings, New York, Queens, Richmond). TLC Taxi trip data for September and October 2013 were downloaded from http://www.andresmh.com/nyctaxitrips. These two months were chosen since they have no major holidays and also had no unusual weather events reported. Each trip record contains pickup and drop-off time and location, passenger count, trip time and distance, and the driver’s taxi permit license.

Average weekday and weekend ridership data for subway stations for the years 2009 to 2014 is available from the Metropolitan Transportation Authority (MTA) Website.
This ridership data together with the subway station geometry is available in ESRI’s Geodatabase format for download (17). The geodatabase contains also the station geometries for three commuter rails (Long Island Rail Road, Metro-North Railroad, and Staten Island Railway) within NYC boundaries as well as a line feature layer with the symbology and geometry of the NYC subway system.

Population and household data at the census tract level from American Community Survey 5-Year Estimates between 2009-2013 at the census tract level were retrieved through ESRI’s Business Analyst. Longitudinal Employer-Household Dynamics (LEHD) Workplace Area Characteristic data at the census block level were used to estimate the number of jobs per census tract. Land use (zoning) vector data were downloaded from MapPLUTO, which is hosted on the NYC Department of City Planning Website.

Data preparation

The study area contains 2166 census tracts. To increase the number of taxi trips per analysis area unit the census tract geometries were grouped into 130 districts using K-means clustering. Socio-economic data (e.g., population, jobs, income) were aggregated accordingly. The taxi trip records as well as the district geometries were imported into a PostgreSQL database with PostGIS extension. The dataset for September and October contained over 29.1 mio. trips. Following trips were removed:

- missing coordinates for pickup or drop-off location
- pickup or drop-off location outside the city boundaries
- trip speed above 100 km/h or below 5 km/h
- trip distance < 300 m

which resulted in 28.2 mio. trips that were retained for further analysis.

Analysis methods

Data analysis is split into two parts. The first part analyzes patterns of taxi trips, whereas the second part explores the relationship between taxi trips and other explanatory variables, such as subway ridership. The first part is of exploratory nature and involves mapping and plotting descriptive statistics related to trip frequency, trip distance, and trip speed, computed for individual districts and/or time periods. It includes also an assessment of the similarity between hourly trip demand patterns on weekdays and Sundays, and uses local statistics (Anselin Local Moran’s I) to identify spatial clusters of districts with high or low similarity values.

The second part uses multiple regression at the district level to determine how transit infrastructure (e.g., subway and bus stations) and subway ridership affect taxi trip demand after controlling for non-network related variables, such as jobs and population. Since observed trip numbers represent a count variable, a negative binomial regression model is used, which can handle overdispersed count outcome variables. Spatial autocorrelation is handled through eigenvector spatial filtering (ESF). A combined measure of taxi trip numbers and subway ridership counts for each district will determine which district shows a tendency towards taxi use. This variable can be analyzed through local statistics (Anselin Local Moran’s I) to identify local clusters of high taxi demand, that is, areas where improved rapid transit infrastructure could be particularly helpful to increase the transit mode share.
PATTERN ANALYSIS OF TAXI TRIPS

Basic trip characteristics
Typical recurring patterns of hourly and daily trip numbers of NYC cabs have been charted and described elsewhere (11), revealing for example that the number of taxis on duty is highest during weekend nights and lowest on Sunday evening. Furthermore, a drop in taxis available can be observed during driver shift change, which takes place around 5am and 5pm. The provided reference plots also the average distance per trip for 2010 taxi trip data, which is similar to the curves plotted in FIGURE 1a and b for the analyzed two-month period in 2013. The graphs indicate that mean trip distance and speed peak in the morning hours, with an earlier peak during weekdays. The similar shape of the distance and speed curves indicate that longer trips are faster, which can be expected given the higher proportion of limited access highways on longer trips (18). A small portion of the morning peak in trip distance can be attributed to the larger proportion of trips to airports (FIGURE 1c) and from airports (FIGURE 1d) in the morning. Trips to and from the two major airports, John F. Kennedy International Airport (JFK) and LaGuardia Airport (LGA), are longer on average (JFK: ~25 km, LGA: ~15 km) than trips to and from other districts (~5 km). Another contributing factor to longer trips in the morning could be the higher share of work related trips (19).

Mean trip distance varies strongly between analyzed districts. FIGURE 2 visualizes for weekdays (Monday through Thursday) the mean travel distance (in km) for trips originating from (FIGURE 2a) and ending in (FIGURE 2b) each mapped district, respectively. Only districts with at least 100 pickups (FIGURE 2a) and drop-offs (FIGURE 2b) are considered. Taxi trip distance increases with distance from the CBD (Midtown and Lower Manhattan), indicating that most trips from and to the CBD are local trips. In fact, the percentage of weekday trips ending in any of three analyzed CBD districts and originating from JFK among all trips ending in the CBD range only between 1.2% and 1.8%. These numbers are somewhat higher for LGA with 2.5% to 2.9%, probably due to its closer proximity to the CBD and its lack of access to the subway system. The maps show also that trips to and from airports are longer than those to and from their surrounding districts.

The role of the CBD for taxi trips is more prominent when considering only trips that begin or end at airports. FIGURE 2c visualizes for weekdays the spatial distribution of drop-offs for trips originating from JFK, and FIGURE 2d shows where passengers were picked up for a taxi ride to JFK. It is apparent that Midtown and lower Manhattan contribute strongly to taxi demand from and to JFK. Furthermore, the high pickup and drop-off numbers for the LGA district indicate that taxi rides occur frequently between JFK and LGA. These trips could stem from passengers who arrive on national flights at LGA and need to connect to international flights departing from JFK. Results shown in FIGURE 2 are comparable to those for Sundays with only minor differences.

Temporal variation of travel speed
Historic or live floating car data are commonly used to determine urban road network travel time and congestion (20, 21). The goal of this analysis is to demonstrate that mean travel speed, as provided in the used dataset, can reproduce a pattern of changed travel speed over time as well. FIGURE 3 maps for different pairs of selected trip departure times the differences in mean travel speed (given in km/h). Only trips up to 8 km long are considered in order to focus on travel speed in and near each mapped district. Furthermore, only districts with at least 30 trip counts for both...
compared travel times are mapped (hence the gaps in the map). FIGURE 3a maps travel speed at 7am subtracted from that at 10pm for weekdays. The positive map values indicate that travel speed is in each district higher at 10pm than at 7am. The largest differences are found around the airports, where demand for taxi trips peaks in morning hours (compare FIGURE 1c and d). In FIGURE 3b, which compares the travel speed between 7am and 10pm on Sundays, most values are negative, suggesting faster travel in the morning than in the later evening. FIGURE 3c compares travel speed at 7am between weekday and Sunday, indicating higher travel speeds on Sunday morning. As opposed to this, FIGURE 3d shows that travel speeds are comparable at 10pm between weekday and Sunday, since mapped differences are close to zero.

**Weekend-weekday correlation in trip frequency**

The comparison of the hourly trip frequency pattern between weekend and weekday can give insight into the functionality of neighborhoods (22). For each district with at least 100 trips on weekday and Sunday, the average number of trips was extracted for each hour on weekday and Sunday. The obtained 24 trip counts for each day were then normalized between 0 and 1. The similarity in taxi demand between weekday and Sunday was then determined through bivariate correlation between the two normalized frequency series. The map in FIGURE 4a reveals positive correlation values for both airports, indicating that the daily demand pattern for taxi trips from each of the airports is similar throughout the week. Some districts in lower Manhattan show smaller correlation values compared to their neighborhood, which could be because of nightlife activities on weekends, or administrative buildings and offices (e.g. banks or city hall) which are closed on weekends.

A local statistic is a descriptive statistic associated with a spatial dataset whose value varies from place to place (23). Local Indicators of Spatial Association (LISA) statistics evaluate the existence of clusters in a spatial arrangement of a variable (24). One example for LISA is Anselin Local Moran’s I, which can identify statistically significant spatial clusters of high or low values and outliers. Applying local Moran’s I to the correlation variable in FIGURE 4a, a High-High (HH) cluster identifies a district with a high correlation value surrounded by other districts with high correlation values, indicating a hot spot. HH clusters can be observed at JFK, around Central park, and in a residential area in the north-east of Brooklyn (FIGURE 4b). A Low-Low (LL) cluster indicates a district with a low correlation value that is surrounded by districts with low values (cold spot). This occurs for one district in the north of the study area. The computation of the levels of significance for clusters identified in FIGURE 4b (which are visualized based on a threshold value of \( p=0.05 \)) has not taken into account multiple testing and spatial dependency. Depending on the correction method fewer clusters might potentially be marked as significant.

The plots in FIGURE 5 show for four selected districts (labeled in FIGURE 4a) the normalized trip counts per hour for weekday and Sunday, sorted by descending order of correlation. JFK (FIGURE 5a) demonstrates the highest correlation among all districts, with a morning peak around 6am for weekday and Sunday, possibly from arriving overnight flights. LGA (FIGURE 5b) shows a steeper taxi demand between 6am and noon on weekdays than on Sundays, possibly due to early arrivals of regional business flights during the week. The higher trip demand between midnight and 6am for Sundays compared to weekdays in the Midtown/Broadway district (FIGURE 5c) could be because of late night entertainment activities offered, which are primarily...
attended on weekend days (e.g., Saturday), leading to an increase in trip frequency after midnight on the next day (i.e., Sunday). Taxi demand picks up faster after 6am on weekdays compared to Sundays due to work trips in the early morning. East Village (FIGURE 5d) is the district with the lowest correlation. It shows high traffic demand after midnight on Sunday (similar to the Midtown/Broadway district), but a lower relative demand afterwards compared to weekdays.

### Relationship Between Taxi Trips and Public Transit

#### Multiple Regression Model

TABLE 1 shows the results of selected (nonspatial) negative binomial regression models (NBMs) for estimating the number of taxi trips from and to districts for weekdays (Monday through Thursday), based on observations from all 130 districts. The models are arranged in a descending order of AIC (Akaike information criterion), where a smaller AIC means a better model fit. Due to bivariate correlations between predictor variables only a subset of predictor variables can be included at a time, otherwise multicollinearity occurs. The coefficients included in the different models are as follows:

- Population, number of jobs
- Socio-economic variables: % of age group, % African American, % Hispanic, % Bachelor of Science degree or higher, mean annual household income, % of households without automobile
- Land use (zoning): % manufacturing, % residential, % commerce, % parks, CBD
- Built environment: Road density (in km/km²)
- Public transportation: Number of metro and train stations, bus stations, metro boardings

Models 1 through 5 (M1-M5) include the number of subway/train stations and bus stations as predictor variables, expressing supply (TABLE 1). The subway/train stations variable is replaced by subway ridership (a demand variable) by the variation of some models shown in the lower part of TABLE 1. All five model estimations (and their variations in the lower half of the table) have in common that population is positively associated with taxi trip counts and airports, which could mean that more people (on airports or elsewhere) increase taxi demand, or that taxis target areas with many potential customers (including airports) more frequently. In none of the models CBD was significant. In M1 the number of subway/train stations is positively associated with taxi trip counts. This is consistent with previous findings where a high correlation between public transit ridership and taxi trips was explained by the direct demand for taxi service from major transit stations (1). Furthermore, the model reveals an increase in the number of bus stops to be associated with a decrease in taxi ride numbers. This could indicate that bus trips in well served areas compete with taxi transportation mode. Ethnicity variables became only significant in combination with land use variables (M2), but did generally not improve the model fit. For all models except for M1, the subway/train station variable becomes non-significant, indicating that much of the variation of taxi ridership associated with rapid transit stations is absorbed by other land user or socio-economic variables. M3 shows that an increase in the percentage of population between 0 and 19 years is associated with a decrease in taxi use. A possible explanation is that many families with children own an automobile and do not have to rely on taxi transport. M3
shows also that parks and residential areas are associated with fewer taxi trips, which could indicate that taxi trips are often used for business trips in more commercial areas. M4 shows that districts with a higher percentage of households without automobile tend to increase taxi trips, as can be expected. Whereas for all other NBMs (M1, M2, M3, M5) the Variance inflation factor (VIF) was < 3.2, indicating that multicollinearity among predictor variables did not pose a problem in the model specifications (25), for M4 the VIF is 4.8, due to the high bivariate correlation of the auto ownership variable with other variables, e.g. number of jobs. Model M5, which includes also education and road density, was found to be the best fitting model while keeping the VIF low. All models have in common that number of bus station is the only significant transit factor associated with taxi trips. This finding can therefore also be considered by smaller municipalities that do not have a subway or train system. A log-likelihood test in all identified NBMs reported overdispersion in the count data (p <0.0001), indicating that a NBM is better suited in modeling taxi trip numbers than a Poisson regression model.

In the lower group of NBMs the job and income variables become non-significant, meaning that subway boarding numbers absorb some of the effect of job numbers on taxi demand. In the lower versions of M2 and M3 subway ridership remains a significant predictor, whereas in M4 and M5 of the lower models it becomes non-significant (hence these modified models are not shown in the table).

Spatially filtered negative binomial model
A fundamental assumption of regression analysis is residual independence. Incorporating spatial data in nonspatial models typically results in residuals that exhibit spatial autocorrelation. Using the regression results of the nonspatial prediction model for the number of weekday taxi trips that reach a district as an example, residual diagnostics reveals significant autocorrelation with a Moran’s I of 0.1120 (p=0.0087). To account for this spatial effect, a spatially filtered NBM is estimated which uses eigenvector spatial filtering (ESF) to model spatial autocorrelation. A detailed description of involved steps is provided in (26, 27). The spatial filter is used as a predictor variable in the negative binomial regression model and expected to explain a considerable part of the variance in the taxi trip crime distribution.

The results in TABLE 2 show that the spatial filter is highly significant (p<0.0001). It eliminates all of the unexplained residual spatial autocorrelation (Moran’s I = -0.0085, p=0.9845). This model satisfies therefore the fundamental assumption of residual independence in count models (28). Comparison with the nonspatial NBM shows a better model fit, as indicated by a lower AIC score. Model comparison shows also that the spatially filtered NBM reduces both the statistical significance as well as the magnitude of all coefficients that were identified as significant in the nonspatial NBM, indicating that violating the assumption of spatial independence in residuals leads to biased coefficients with inflated levels of significance. Compared to the nonspatial NBM, in the spatially filtered NBM the number of bus stops is no longer a significant predictor. Some land use variables (percent residential area, road density), airport, as well population (total and percent between 0 and 19 years) remain the only variables with significant coefficients, besides the spatial filter. This means that based on global regression models there is no clear direction of the effect of transit supply, e.g. subway, train and bus stations, on taxi trip demand. Instead the direction of the effects, if present, may vary over space and time, which can be explored through
independent local spatial models (4), separate models by the hour (9), or by adding interaction terms with location and time of day in the regression model. In this regard, spatially-filtered NBMs serve as a reminder that results of nonspatial regression models should be interpreted with care, especially when spatial autocorrelation is present in residuals.

**Joint spatial pattern of taxi rides and subway use**

The following analysis aims to identify districts of the city with an increased tendency towards taxi use compared to metro use, based on observed ridership data. Previous regression analysis did not identify a clear relationship between subway supply and taxi demand (see TABLE 1 and TABLE 2). That is, presence of metro stations will in some cases increase the demand for taxi trips, e.g. for last mile trips. A city may also have districts where taxi demand is high because of a lack of rapid transit infrastructure and no other realistic transportation alternatives than taxi (e.g. due to low car ownership of travelers). Such areas have a potential to increase rapid transit ridership, e.g. through infrastructure improvements or policy changes. Although the NYC transit network operates besides the subway also three commuter rails, the following analysis considers subway ridership only and is therefore limited to 101 districts that contain subway stations or are located between them (FIGURE 6).

To describe the tendency of a district towards taxi or metro use numerically a modemix variable $M$ is introduced and computed for each district $d$ as

$$M_d = \frac{\text{Taxitrips}_d}{\text{MaxTaxirides}} - \frac{\text{Subwayridership}_d}{\text{MaxSubwayridership}}$$

where $\text{Taxitrips}$ and $\text{Subwayridership}$ are the total taxi trip count and subway ridership in district $d$ for a given time period, and $\text{MaxTaxirides}$ and $\text{MaxSubwayridership}$ are the maximum values of $\text{Taxitrips}$ and $\text{Subwayridership}$, respectively, among all districts in the entire study area. $M$ ranges between -1 and 1. An $M$ value below zero indicates a travel behavioral tendency towards subway use, and an $M$ value above zero indicates a tendency towards taxi use. Districts without access to subway but observed taxi trips will always receive an $M$ value above zero since the second term in the equation is zero. Among this group of districts, those with a large number of taxi rides, e.g. triggered by many jobs or a large population, will tend to have higher $M$ values. For such districts, providing rapid transit access could be an effective way to reduce the number of taxi rides and substitute them with transit rides.

FIGURE 6a maps for each district the number of taxi trips that pick up passengers on a weekday, and FIGURE 6b shows the total subway ridership on a weekday. JFK does not have a subway station but an AirTrain system which connects to Jamaica and Howard Beach subway stations. Using the daily number of AirTrain trips, the known split of AirTrain users boarding at Jamaica and Howard Beach station (approximately 60/40), and assuming a 50/50 split of JFK Train users to continue their journey in the Long Island Road System and the subway system, respectively, the corresponding number of daily AirTrain ridership was shifted from Jamaica and Howard Beach station to the JFK analysis district. FIGURE 6c visualizes the modemix variable, revealing higher values, i.e. a tendency towards taxi use, for the two airports and large portions of Manhattan. FIGURE 6d maps the result of the Anselin Local Moran’s I procedure on the modemix variable (without considering multiple testing and spatial dependency). The High-Low (HL) category indicates a high value district surrounded by low value districts, and the Low-High (LH) category indicates the opposite. Both the HL and LH case indicate a spatial outlier. A region of HH clusters, i.e. a hot spot of taxi demand, can be clearly discerned in the Manhattan area.
high taxi demand relative to metro use could be caused by a limited capacity of the subway system (e.g. crowded trains and stations) during rush hour. Another potential explanation is a reduced sensitivity of travelers to higher taxi trip fares (compared to subway) in these districts, given high household income (e.g. around Central Park), and well paid jobs in these districts. Yet another possible reason is that the CBD has the lowest trip distance in NYC (see FIGURE 2a, b), which keeps taxi fares generally affordable. Given that Manhattan has already a dense network of subway lines, adding more subway infrastructure will most likely not reduce taxi ridership. The situation is different for the marked High-Low outlier around the LGA airport where the higher modetrix value of the LGA district compared to its surrounding districts could be attributed to missing rapid transit infrastructure at LGA. As opposed to some of its surrounding districts LaGuardia offers currently no subway or rail service, but is only served by several bus lines. Because of this deficiency, a new LGA AirTrain system is currently proposed, which will be a 1.5-mile-long (2.4 km) people mover system connecting with the subway and Long Island Rail Road in a similar manner to AirTrain JFK (29). LISA results in FIGURE 6d support this endeavor by showing the outlier with an unusually high taxi mode share in the LGA area.

SUMMARY AND FUTURE WORK
This study explored various aspects of travel behavioral analysis in an urban environment using a freely available dataset of taxi trip records for NYC, as well as subway ridership data. Aggregate analysis for the complete study area showed a clear difference in taxi demand throughout the day when comparing weekdays and Sundays patterns, which can be attributed to different types of activities conducted on weekdays and Sundays. A more refined picture of the variation of activities between different days and regions can be drawn by correlating hourly trip count values on weekdays and Sundays for each district. This analysis can be used to distinguish between different functional areas, such as airports, residential areas, business districts, or entertainment districts. Combining taxi trip counts with subway ridership data was shown to provide useful information about areas that might be currently underserved by public transport, and where observed high taxi trip numbers indicate a higher demand for alternative transportation, such as rapid transit. To draw more detailed conclusions, the analysis would need to include heavy rail and bus services as well, which is deferred to future work. Some of the nonspatial binomial regression models presented reveal that subway and train stations are associated with an increase in taxi trips, confirming results from previous studies (1, 4). This effect, becomes, however statistically non-significant when using a spatially filtered binomial regression model that explicitly models spatial autocorrelation. This shows that this relationship does not hold throughout the study area, but may have to be addressed in local regression models. Bus stations tend to be associated with lower taxi demand. However, also this effect disappears in the global model when taking into account spatial autocorrelation.

ACKNOWLEDGEMENTS
I want to thank Sreten Cvetojevic for his help with the management of the PostgreSQL database which was used for data analysis in this study.
REFERENCES


List of Table Titles and Figure Captions

TABLE 1 Negative Binomial Regression Estimates for Weekday Taxi Trips from Districts

TABLE 2 Estimation Results for a Nonspatial and Spatially Filtered Negative Binomial Model for Weekday Taxi Trips to Districts.

FIGURE 1 Trip distance (a), trip speed (b), and share of trips to airports (c) and from airports (d) over a day.

FIGURE 2 Weekday distance (in km) for trips originating (a) and ending (b) in mapped districts, and number of trip destinations (c) and origins (d) for trips from and to JFK.

FIGURE 3 Difference in mean travel speed for different times of the day (a, b) and different days (c, d). Numbers are given in km/h.

FIGURE 4 Correlation in daily trip frequencies between weekday and Sunday (a), and Anselin Local Moran’s I (b).

FIGURE 5 Correlation of daily trip frequencies between weekday and Sunday for selected districts.

FIGURE 6 Average number of taxi trips (a) and subway boardings (b) on a weekday, the resulting modemix (c), and Anselin Local Moran’s I (d).
**TABLE 1 Negative Binomial Regression Estimates for Weekday Taxi Trips from Districts**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>5.268**</td>
<td>5.782**</td>
<td>14.452**</td>
<td>4.144**</td>
<td>11.148**</td>
</tr>
<tr>
<td>% 0-19 years</td>
<td>-0.213**</td>
<td>-0.114**</td>
<td>0.017**</td>
<td></td>
<td>-0.254**</td>
</tr>
<tr>
<td>% Afr. American</td>
<td></td>
<td>0.017**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% BS</td>
<td></td>
<td></td>
<td></td>
<td>0.051*</td>
<td></td>
</tr>
<tr>
<td>Jobs</td>
<td>1.461E-05**</td>
<td>1.040E-05**</td>
<td>8.480E-06**</td>
<td>7.528E-06*</td>
<td>5.995E-06*</td>
</tr>
<tr>
<td>Income</td>
<td>1.684E-05*</td>
<td></td>
<td>2.534E-05**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subway/train station</td>
<td>0.137**</td>
<td>0.058</td>
<td>-0.035</td>
<td>-0.058</td>
<td>0.017</td>
</tr>
<tr>
<td>Bus stops</td>
<td>-0.027**</td>
<td>-0.026**</td>
<td>-0.024**</td>
<td>-0.013*</td>
<td>-0.016**</td>
</tr>
<tr>
<td>Airport</td>
<td>7.190**</td>
<td>7.801**</td>
<td>5.727**</td>
<td>8.140**</td>
<td>7.307**</td>
</tr>
<tr>
<td>% Residential</td>
<td>-0.044**</td>
<td>-0.045**</td>
<td>-0.060**</td>
<td>-0.062**</td>
<td></td>
</tr>
<tr>
<td>% Park</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% HH without car</td>
<td></td>
<td></td>
<td></td>
<td>0.100**</td>
<td></td>
</tr>
<tr>
<td>Road density</td>
<td>0.222**</td>
<td></td>
<td></td>
<td>0.187**</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>2265.6</td>
<td>2226.2</td>
<td>2196.4</td>
<td>2184.7</td>
<td>2172.4</td>
</tr>
</tbody>
</table>

| Intercept                  | 5.603**  | 5.786**  | 14.339** |          |
| Population                 | 5.184E-05** | 5.732E-05** | 6.646E-05** |          |
| % 0-19 years               |          | -0.208** |          |          |
| % Afr. American            | -0.017** |          |          |          |
| % BS                       |          |          |          |          |
| Jobs                       | 1.491E-05 |          |          |          |
| Income                     |          |          |          |          |
| Subway boardings           | 2.731E-05** | 1.314E-05** | 8.638E-06* |          |
| Bus stops                  | -0.024** | -0.022** | -0.022** |          |
| Airport                    | 7.179**  | 7.735**  | 5.806**  |          |
| % Residential              | -0.045** | -0.045** |          |          |
| % Park                     |          |          |          |          |
| Road density               | 0.228**  |          |          |          |
| AIC                        | 2262.5   | 2222.8   | 2197.0   |          |

Note: ** p<0.01, * p<0.05, + p<0.1 (statistical trend)
TABLE 2 Estimation Results for a Nonspatial and Spatially Filtered Negative Binomial Model for Weekday Taxi Trips to Districts.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Nonspatial Neg. Binomial Model</th>
<th>Spatially filtered Neg. Binomial Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>Standard error</td>
</tr>
<tr>
<td>Intercept</td>
<td>11.273**</td>
<td>0.859</td>
</tr>
<tr>
<td>Population</td>
<td>4.429E-05**</td>
<td>3.631E-06</td>
</tr>
<tr>
<td>% 0-19 years</td>
<td>-0.184**</td>
<td>1.970E-02</td>
</tr>
<tr>
<td>% BS</td>
<td>3.259E-02*</td>
<td>1.306E-02</td>
</tr>
<tr>
<td>Sub/train sta.</td>
<td>3.317E-02</td>
<td>3.002E-02</td>
</tr>
<tr>
<td>Bus stops</td>
<td>-6.048E-03**</td>
<td>2.173E-03</td>
</tr>
<tr>
<td>Airport</td>
<td>4.465**</td>
<td>0.736</td>
</tr>
<tr>
<td>% Residential</td>
<td>-2.922E-02**</td>
<td>5.435E-03</td>
</tr>
<tr>
<td>% Park</td>
<td>-2.014E-02*</td>
<td>8.098E-03</td>
</tr>
<tr>
<td>Road density</td>
<td>8.664E-02**</td>
<td>2.550E-02</td>
</tr>
<tr>
<td>Spatial filter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>2626.1</td>
<td></td>
</tr>
</tbody>
</table>

Note: ** p<0.01, * p<0.05, + p<0.1 (statistical trend)
FIGURE 1 Trip distance (a), trip speed (b), and share of trips to airports (c) and from airports (d) over a day.
FIGURE 2 Weekday distance (in km) for trips originating (a) and ending (b) in mapped districts, and number of trip destinations (c) and origins (d) for trips from and to JFK.
FIGURE 3 Difference in mean travel speed for different times of the day (a, b) and different days (c, d). Numbers are given in km/h.
FIGURE 4 Correlation in daily trip frequencies between weekday and Sunday (a), and Anselin Local Moran’s I (b).
(a) JFK; \( r = 0.95 \)

(b) LGA; \( r = 0.86 \)

(c) Midtown/Broadway; \( r = 0.30 \)

(d) East Village; \( r = -0.37 \)

FIGURE 5 Correlation of daily trip frequencies between weekday and Sunday for selected districts.
FIGURE 6 Average number of taxi trips (a) and subway boardings (b) on a weekday, the resulting modmix (c), and Anselin Local Moran’s I (d).