Paper No. 14-4217

Modeling Taxi Trip Demand by Time of Day in New York City

Ci Yang, M.S. (corresponding author)
Graduate Research Assistant
Department of Civil and Environmental Engineering
Rutgers, The State University of New Jersey
96 Frelinghuysen Road, Piscataway, NJ 08854
Tel: (631) 880-1028
Email: jessie_yang06@yahoo.com

Eric J. Gonzales, Ph.D.
Assistant Professor
Department of Civil and Environmental Engineering
University of Massachusetts Amherst
130 Natural Resources Road, Amherst, MA 01003
Tel: (413) 545-0685, Fax: (413) 545-9569
Email: gonzales@umass.edu

Word Count

Abstract:	203
Text:	5421
Figures/Tables:	6 × 250 = 1500
Total:	7124

Paper Submitted for Publication in the *Transportation Research Record*
Submission Date: Mar 15, 2014

ABSTRACT
Identifying the factors that influence taxi demand is very important for understanding where and when people use taxis. A large set of Global Positioning System (GPS) data from New York City (NYC) taxis is used along with demographic, socioeconomic, and employment data to identify the factors that drive taxi demand. A technique is developed to measure and map transit accessibility based on the Transit Access Time (TAT) to understand the relationship between taxi use and transit service. The taxi data is categorized by pick-ups and drop-offs at different times of day. A multiple linear regression model is estimated for each hour of the day to model pick-ups and another to model drop-offs. Six important explanatory variables are identified that influence taxi trips: population, education, age, income, TAT, and employment. The influence of these factors on taxi pick-ups and drop-offs changes at different times of the day. The number of jobs in each industry sector is an indication of the types of economic activities occurring at a location, and in some sectors the number of jobs is strongly associated with taxi use. This study demonstrates the temporal and spatial variation of taxi demand and shows how transit accessibility and other factors affect it.

INTRODUCTION

Taxis in New York City (NYC) carry 172 million trips annually (11% of all travel), making the cabs an important transport mode in the city (*1*). All NYC taxis are regulated by the Taxi and Limousine Commission, which issues medallions and sets fare rules although cab drivers choose where to circulate to pick up passengers. In order to effectively plan and manage the taxi fleet, it is necessary to understand what factors drive taxi demand, how taxi use is related to the availability of public transit, and how these patterns vary over space and time. A trip generation model that relates taxi demand to observable characteristics of a neighborhood (e.g., demographics, employment, and transit accessibility) is useful for planners and policy makers to manage taxi services effectively.

Trip generation models are used to predict the total number of trips that originate or terminate in a Transportation Analysis Zone (TAZ), and this constitutes the first step of a travel demand forecast (*2,3*). These models relate the total number of trips produced in a TAZ to a variety of factors related to the TAZ and transportation modes available (*2,4,5,6*):

- Level Of Service (LOS) of the mode;
- accessibility of the mode;
- demographics of the TAZ (e.g., population, race);
- socioeconomics of the TAZ (e.g., income, education);
- other characteristics of the TAZ (e.g., land area);
- land use in the TAZ.

Three methods are commonly used to model trip production: rate method (*7*), cross classification (*2,5*), and regression (*3,8*). The rate method is used for traffic impact analysis on non-residential trip generation, which does not consider characteristics such as household size, income, and auto ownership. A cross classification model cross-tabulates average trip making rates with two or more variables, revealing important factors without assuming that the relationship between demands and explanatory variables follows a specific functional form or that there is independence between these factors. The regression method produces a Maximum Likelihood Estimate (MLE) for the coefficient of each explanatory variable in a model that implies a functional relationship between the explanatory variables and the dependent variable.

Regression is a widely used statistical method for exploring the relationship between response variables and explanatory variables with various approaches for validating the model. If enough information is available, trip generation based on regression models can be very useful to forecast travel demands in each TAZ of an urban transportation system (*2,3*). A large dataset with sufficiently detailed information about travel and TAZ characteristics is necessary to model trip generation across a large geographic area using regression.

In this study, taxi trip information from ten months of complete Global Positioning System (GPS) data is related with transit information in NYC. These spatially and temporally classified data are the response variables to be modeled. Possible explanatory variables that relate to taxi demand include aggregate data at the level of census tracts including population, household income, education, total employment, and types of jobs. Other factors that potentially influence taxi demand include the LOS and the accessibility of transit at each TAZ (*2,3*), but these require that detailed transit schedule information be cleaned and compiled before inclusion in the model. Detail taxi and transit data that has both spatial and temporal components allows for investigation of how the factors that drive taxi use change at different times of the day.

The paper is organized as follows: first, the literature on factors that have been found to influence trip generation is reviewed; a section on data that provides a description of the taxi

GPS data and the explanatory variables in this study follows; then a novel technique to calculate transit accessibility and a multiple linear regression model to identify influential factors is presented. The model results are presented following the methodology, and conclusions are discussed.

LITERATURE REVIEW

There are few studies using large taxi GPS datasets to model taxi trip generation. Schaller (*6*) presents an analysis of the number of taxicabs in 118 U.S. cities using multiple linear regression models. The factors influencing the size of a city's taxi fleet include population, employment, use of complements to taxi cabs (e.g., transit), cost of taxis, and taxi service quality. However, the model predicts the quantity of taxi cabs instead of the number of taxi trips generated. The number of workers commuting by subway, the number of households with no vehicles available, and the number of airport taxi trips have significant explanatory power for the number of cabs in operation. Mousavi et al. (*8*) stated that household structure, age, gender, marital status, income, employment, car ownership, population density, and distance to transit are the most influential variables on trip generation for all modes.

Taxi demand may be closely related to transit accessibility in NYC, because taxis and transit are both provide transportation service to the public. The factors that influence transit use may also have an effect on taxi trip generation, but the tendency to choose transit versus taxi may also be affected by the accessibility of transit near trip origins and destinations. For this reason, factors related to transit and vehicle modal split could be included in models for taxi demand. Racca and Ratledge (*4*) presents a comprehensive list of possible factors that are used for mode choice modeling including transit LOS, accessibility, land use, demographics, and characteristics of the trips. That study shows that high transit service is focused at locations with high employment and population densities in the city of Wilmington, Delaware. The analysis on mode split versus mean age and time-of-day indicate that these variables affect the modes that people choose, and this means that they may also relate to taxi trip generation. Corpuz (*10*) shows that socio-economic characteristics and time-of-day have influenced people's choices between private vehicles and public transportation. Workers and households with higher incomes are more likely to use cars over public transit in that time-of-day analysis. The train and the bus are more likely to be picked during morning and late afternoon peaks, because people want to avoid the time and the cost of driving in congestion (*10*).

Characteristics of the trip (e.g., travel purpose) and characteristics of the traveler (e.g., age) have been identified as influential factors affecting the trips generated by different travel modes (*2,5,10,11*). Trips to residential areas and non-residential areas (*12*) and trips for business and non-business purposes (*2*) are analyzed separately in most studies. A number of studies have been conducted about the generation of airport trips (*11*) and travel to schools (*3,13*). Researchers have also studied trips generated by elderly people, because their needs and behavior have some distinct differences from other population groups (*11,14*).

Without detailed information about the taxi trip purpose or the characteristics of the specific person making each trip, the methods in this paper make use the characteristics of the places where taxi trips start and end to gain insights about the demographic and land use factors that are most associated with taxi trip making. This paper focuses on the characteristics of the people who live and work in these places in order to develop models for taxi trip generation.

DATA

Trip generation models require comprehensive sets of data for explanatory variables in order to identify the most influential factors on taxi trip generation. The database of taxi trips has complete information of 147 million taxi trips made between February 1, 2010 and November 28, 2010, including temporal and spatial information acquired by GPS (taxi pick-up and drop-off date, time, location), fare (including tolls, tip, total fare paid), and distance travelled. The taxi data for pick-up and drop-off locations are aggregated by hour of the day in a similar manner to the way that taxis were used as traffic probes by time of day in Yazici et al. (*9*). The distribution of pick-ups (origins) and drop-offs (destinations) are considered separately because they are clustered differently in time and space. Thus separate models are developed to understand these two trip ends. Since census tracts are the TAZs in this study, all data are grouped by census tract so that the response variable and explanatory variables are aggregated at the same spatial resolution.

The sources of data for the explanatory factors considered in this study include:

- transit LOS based on NYC subway schedules available from Google Transit Feed Data in the format of General Transit Feed Specification (GTFS);
- demographics data for each census tract is available from the U.S. Census 2010, including total population, population categorized by age, and population categorized by race;
- socioeconomic data is available from the American Community Survey 5-year estimate of education and income;
- employment data by census tract, including categorization by age, earnings, type, race, ethnicity, educational attainment, and sex is available for NYC from 2010 Workplace Area Characteristic (WAC) data from U.S. Census Bureau;
- geographic data including relevant shape files (i.e., rivers, roads, county, census tract), and land area.

These data are explanatory variables that are included in the model (e.g., the response variable is produced by using the taxi data and census tract geographic information). The population density and employment density in 2010 is calculated for all 2,167 census tracts in NYC. Figure 1 shows that the population density and employment density are concentrated in Manhattan. Compared with the taxi demand information in Figure 2, it is clear that pick-up demand is concentrated in Manhattan, northern Brooklyn, and the west and north sides of Queens while the drop-off demand is more spread over four of the boroughs: Manhattan, Brooklyn, Queens, and Bronx.

Some census tracts consisting of cemeteries, parks, or islands do not have employment associated with them, so the WAC employment data covers 2,143 census tracts. Census tracts that variables are lacking certain required information are excluded from the linear model analysis. Ultimately, 116 out of 2,167 census tracts (5%) were omitted from the analysis, because there was insufficient population or employment in those few regions to create a useful data point.

METHODOLOGY

There are two important methodological contributions of this study. The first is the development of a novel transit accessibility measure based on the time to access and wait for transit. This requires processing raw transit schedule information to determine how much time it takes person at a specific location and time of day to access the public transit system. The second is the development of a hybrid cross-classification/regression model for estimating taxi trip generation.

The taxi data is cross-classified by pick-up and drop-off and aggregated by hour of the day. Within each classification, a multiple linear regression model is estimated to identify the factors that influence taxi demand.

Transit Access Time

Transit LOS and accessibility must be quantified in order to be used as an explanatory variable to model taxi. A new measure is developed that combines the estimated walking time a person must spend to access the nearest station (transit accessibility) and the estimated time that person will wait for transit service (transit LOS). This measure is the Transit Access Time (TAT), and it represents the minimum expected time for a person at a specific location and time-of-day to walk to, wait for, and board a transit vehicle. For a walking speed of 3.1 mph (5.0 kph) (*15*), the transit access time in minutes is:

$$\text{TAT} = \frac{60D}{v_w} + \frac{60}{f} \quad (1)$$

where f is the frequency of subway dispatches per hour at the nearest station, D is the distance to the nearest station (mi), and v_w is the walking speed (mph).

The minimum TAT is calculated at each location by the following steps. First, the transit schedule in GTFS provides the number of transit departures (i.e., frequency) in each hour at each station. The waiting time depends on the frequency based on the second term of Equation 1, and it is calculated separately for each hour of the day to account for variations in the schedule. Then, a fine grid is imposed on the study area with cells measuring 250 meters (820 feet) square, which is small enough that the walking time to cross each cell is less than 1 minute. Each cell is characterized by the location of its centroid, and a TAT will be calculated for each cell. A modified K-nearest neighbor algorithm is implemented by calculating the minimum TAT from the k nearest transit stations by screening distance and waiting time to all transit stations from the centroid.

People are assumed to be well-informed about transit schedules and to choose the nearby station that minimizes the sum of their walking and waiting time. Thus, the TAT is a metric of transit accessibility that is independent of specific origin-destination demand patterns. For simplicity, the method looks only at the closest access from each location (cell centroid) to the nearest subway departure, in space and time, anywhere in the system. The minimum TAT is calculated for each cell in NYC at each hour of the day, and this is used to quantify transit accessibility in the city with spatial resolution of 250 meters (820 feet) and temporal resolution of an hour.

Once the minimum TAT for each census tract is determined by averaging the values across the cells included within the census tract. This is provides a better TAT measure than simply calculating from census tract centroids, because a large census tract may have a centroid near a transit station but lots of land that has relatively low accessibility. The TAT is calculated for different times of day for each census tract using only the subway data in this study, because the complete GTFS bus schedule data is incomplete (e.g., bus data for Queens are not available).

Visualizing TAT and Taxi Demand

Figure 2 shows the TAT for subways at 12:00 a.m. (midnight) and 5:00 p.m. (afternoon) along with both taxi pick-ups and drop-offs per capita in the same hours. The map of TAT shows that there is greater transit accessibility in Manhattan and along the subway routes than in other parts of the city, which is expected based on the spatial coverage of the subway network. The transit

accessibility is also generally greater at 5 p.m. than at 12 a.m., because services operate more frequently during the peak hours than late at night. Figure 2 suggests that the pick-ups and drop-offs per capita are higher where the TAT is lower (i.e., transit is more accessible), which is a negative correlation between TAT and taxi use.

It is necessary to separate trips by the hour of the day, because the distribution of activities in NYC changes with time. There are also differences between the rates of taxi pick-ups per capita at 5 p.m. and at 12 a.m. For example, there are more taxi pick-ups at Jamaica at 5 p.m. than that at 12 a.m., which could result from people getting off the subway at Jamaica and then taking a taxi complete a trip home from work. In some areas of lower Manhattan there are more pick-ups at 12 a.m. than at 5 p.m., which indicates concentrations of nightlife.

The drop-offs per capita show big differences between 5 p.m. and 12 a.m. as well. For example, there are more drop-offs per capita at some popular locations such as Penn Station, Grand Central Station, and Flushing at 5 p.m. than at 12 a.m., which is consistent with the fact that these are busy transit hubs that used by commuters. Although the total amount of travel activity in the city is lower at midnight than at 5p.m., many areas of the outer boroughs actually see a greater rate of drop-offs in the late night hours. This suggests that people use taxis more often to travel to outlying neighborhoods when it is dark and transit services are less frequent. There appears to be a consistent trend at all times of day that pick-ups are more concentrated around transit hubs and central areas, whereas drop-offs are more dispersed around the city. Clearly, trip making behavior by taxis is asymmetric.

The mapping of TAT and taxi demand provide a visualization of their relationship and help provide intuition about why such a relationship exists. With the hourly data for TAT, taxi pick-ups, taxi drop-offs, and all other demographic and socioeconomic information, visual inspection of the maps is interesting but insufficient for determining the quantitative relationship between the explanatory variables and the taxi demand. A multiple linear regression model is introduced in the next section in order to achieve this objective.

Taxi demand model

Linear models have been broadly applied to trip generation (*3,8*). The idea behind multiple linear regression modeling is to explore the relationship between the dependent variable and independent variables with the assumption that this relationship is linear as follows

$$Y = \sum_{i=0}^{n} \beta_i X_i + \varepsilon \quad (2)$$

where Y is the number of taxi trips generated in a TAZ (response variable), X_i is one of n independent variables, X_0 is the intercept, β_i is the coefficient corresponding to X_i, and ε is the error representing the difference between the modeled and observed number of taxi trips.

The response variable in the model is the number of pick-ups or drop-offs generated in each census tract by hour of the day from the 10-month taxi GPS data in NYC. The full list of explanatory variables considered in the initial model are listed in Table 1. Using least squares estimation (i.e., MLE) coefficients are estimated for each explanatory variable by minimizing the mean squared error between the modeled Y and observed Y. The goal is select a set of explanatory variables that results in low model error and in which each explanatory variable has a statistically significant coefficient. There are many methodological and statistical criteria for selecting important variables. For example, step-wise selection and best subset regression are two methods for comparing model specifications in order to identify the best set of explanatory

variables to include in the final model. The following steps describe several procedures used to select important variables in this study.

Check Correlation Coefficients

An analysis of the correlation coefficients among the response variable and all explanatory variables shows how closely each pair of variables vary with each other. A correlation coefficient that is greater than 0.5 or less than -0.5 is considered strong in this analysis. The strong correlation between an explanatory variable and the response variable could indicate the explanatory variable is important. Strongly correlation among explanatory variables leads to multicollinearity in the model, because it is not possible to identify which factor has the more significant statistical relationship with the response variable. The Variance Inflation Factor (VIF) quantifies the severity of multicollinearity in an ordinary least squares regression by measuring how much the variance of an estimated regression coefficient increases due of multicollinearity (*16,17*). Each indicator has a VIF value to indicate the degree of multicollinearity, and a large value indicates that a variable needs to be either removed or replaced. A common rule of thumb is that if the VIF of each factor is larger than 5, then multicollinearity is high (*16,17*).

Step-wise Selection

Step-wise selection (or forward and backward selection) is a method of variable selection by adding or eliminating one variable at a time. The best model is chosen by seeking the model with the lowest value of Akaike Information Criterion (AIC) and smaller Residual Sum of Squares (RSS). AIC is a measure of the complexity of the model, and it is a function of maximum likelihood and the number of parameters included in the model. A smaller AIC value indicates a better goodness of fit (*18,19*). The AIC value is especially useful when comparing models with a large number of explanatory variables. The step-wise method involves ranking the importance of each factor by listing the AIC values would result from removing it. Then, the least relevant factors can be eliminated one by one until a suitable model is specified.

Best Subsets Regression

Best subsets regression (a.k.a. complete subset regression) is a method to select the best subset of predictors among all the possible combinations of predictors (2^k combinations if there are k predictors in the initial model) (*20,21,22*). There are several metrics for comparing model performance:

- R squared (R^2) is the coefficient of determination that quantifies the variance in the model error, and it is also an indicator of how well the model fits the data points.
- Adjusted R squared ($AdjR^2$) is similar to R^2 but incorporates a penalty for the number of extra explanatory variables added to the model; a higher $AdjR^2$ is better.
- Bayesian Information Criterion (BIC), which is similar to AIC, is a function of the maximized value of the likelihood function and the number of variables included in the model. The difference from AIC is that the penalty term for the number of variables included in the model is larger in BIC than in AIC (*18,19*). For both metrics, lower values are an indication of a better model.
- Mallow's C_p assesses overfitting of the model, and a desirable model has C_p close to the number of explanatory variables, p (*23*).

The best subset method works well to refine the selection of explanatory variables from the important factors that are already identified. It is very useful for modeling the same major pick-

ups and drop-offs at different times of day based on the same set explanatory variables, because trips at different times of day could be associated with different explanatory factors.

It is very difficult to achieve an R^2 greater than 0.8 in most trip generation studies, because there are many things affecting the response variable, and we seek the simplest possible model to gain insights for transportation planning. There have been some studies by transportation planners on regional growth (*24*) and trip generation (*25*) using linear regression and achieving very low R^2 or adjusted $AdjR^2$ (much less than 0.5 sometimes less than 0.1), however the value of these models is not in the final estimate of the response variable but in identifying statistically significant explanatory variables that help us understand what drives demand. The goal of this study is to identify the relationships between taxi demand and important socioeconomic and land use factors at different times of day and at different locations. Therefore the models are developed not only based on R^2 but also on other criteria used to select an appropriate model. In order to use the fewest number of variables for the model, the most statistically significant explanatory variables are identified by the t-statistic or p-value (p-value<0.05 is significant at 95% confidence level).

RESULTS AND DISCUSSION

The methodology presented in the previous section was used to identify several influential factors from the initial full model using step-wise selection based on AIC values and RSS: TAT, total population, median age, three types of income, total jobs, jobs by type, and jobs by sex, which are listed in Table 1 with a '*' sign.

The correlation coefficient is checked to remove factors that are too closely related to each other in selecting major factors for the second model. Since median household income, mean family income, and per capita income are highly correlated with each other, only one should be included in each model to avoid multicollinearity. Due to better performance of the model with per capita income and the higher correlation coefficient with the response variables, per capita income has been selected. Similarly, jobs by type or jobs by sex are closely related to total jobs. In this case, total jobs, which are an indication of total economic activity in an area, is chosen for the second model with major factors listed in Table 1. To prevent multicollinearity, only one factor or category among two or more correlated factors is included.

Models with and without the intercept are estimated for pick-ups and drop-offs for each hour of day in NYC. In most of the models the intercept is not significant, and it is intuitive that if a census tract has no population and no jobs, then there are likely to be no trips as well. The coefficients of the other explanatory variables are very similar whether or not the intercept is included in the model. Therefore, the intercept is removed from the models formulated in this study. The results, including the 6 major variables for each time of the day, are presented in Table 2. All coefficients are significant (p-value < 0.05) unless labeled "*" for Tables 2, 3 and 4.

The interpretation of the trip generation results for both pick-ups and drop-offs is useful for transportation planning and regulation of taxi services. The magnitude and sign of the coefficient for each explanatory variable indicates how much taxi demand will increase (for positive coefficients) or decrease (for negative coefficients) as the explanatory variables increase by one unit. For example, the coefficient of 'TotJob' is 0.32 for pick-ups at 12 a.m. in NYC (Table 2) indicating that an increase of one job in a census tract is associated with an average increase of 0.32 taxi trips in the 12 a.m. hour over a 10-month period. Similarly, there is an average decrease of 36 taxi trips at the same hour over a 10-month period as transit access time (TAT) increases by one minute, which provides intuition about how dramatically taxi demand changes with the availability and accessibility of transit service.

The errors of the trip generation model (i.e., difference between observed and modeled taxi demand) provide information on when and where taxi demand is underestimated or overestimated. This gives some idea of where and when more taxi use would be expected than actually occurs, based on city-wide trends, so the information can be useful for planning locations of taxi stands or providing incentives for cab drivers to operate over certain times of day and in certain parts of the city. At locations where the model estimates higher taxi use than is actually realized, it is possible that there is a latent demand that goes underserved because there are simply not enough taxis circulating at the specific location and time to carry as many passengers as would like to use taxis.

The results show that population, education, income, and total jobs positively influence both taxi pick-ups and drop-offs in NYC. This is expected, because high total population and high total number of jobs are indicators of places with high human activity and where people are more likely to be traveling by any mode, including taxi. However, median age and TAT negatively affect the trip-making by taxis. This shows that younger people are more likely to take taxis. The results also show that taxi demand is high where transit is more accessible (TAT is small). It is not clear from the available data whether the relationship between taxis and transit is competitive or complementary. Thus, it cannot be concluded whether the convenience of transit service in an area causes high taxi demand because people use taxis to complement transit or if the large number of taxi trips are associated with high levels of activity that also happen to be where high levels of transit service are provided. The reality is likely that taxis and transit are sometimes operating in competition and other times as complements, because both modes follow and influence the levels of activity in neighborhoods across the city.

The distribution of coefficients at different times of day also sheds light on how those factors influence the number of taxi trips (Table 2). For example, the total number of jobs has a higher influence on taxi demand from 7 a.m. to 6 p.m., which indicates that extra taxi demand during this period in NYC is likely caused by people going to and from work or work-related activities. The coefficients for TAT values from 8 a.m. to 11 p.m. show increased taxi trips associated with good transit accessibility (short TAT) during all but the late and overnight hours, so it is possible that many of the trips are being made to or from transit facilities, enabling taxis to complement transit service. It is also possible that the places that have good transit service are also desirable for taxi use for other reasons. For example, it might be easier to hail a cab on busy streets in Manhattan under which the busiest subway lines also run.

Another interesting observation from the stepwise modeling is that some of the variables in the category of jobs by type are very influential in the linear model performance, especially for the pick-ups and drop-offs in Manhattan as listed in Tables 3 and 4. TAT loses its influence for drop-off trips in Manhattan compared to when total jobs was used. Factors related to job types seem to play key roles in generating the taxi trips in Manhattan, some influential industry sectors include jobs in retails, accommodation and food service, and health care (see Tables 3 and 4).

From 11 p.m. to 8 a.m., it looks like the drop-off taxi demand is not significantly related to income while from 9 a.m. to 10 p.m. it is. This indicates that people are taking taxis in the evening no matter how much money they earn, however, in the daytime wealthy people are more likely to take taxis, perhaps because more competitive affordable travel modes are available during the day. Similar situations are also observed for taxi pick-up demand except that the time period is slightly earlier.

People tend to take taxis to places with retail activities from 8 a.m. to 4 p.m. (Table 3) and taxi trips away from these places from 12 p.m. to 11 p.m. These retail related activities could be working to sell goods, shopping to purchase goods, or meeting with other people. Unfortunately, without data about individual trip purposes, it is not possible to say what exactly each traveler did in the census tract, but the high correlation with retail activities shows the importance of retail land use and employment in determining taxi demand.

Accommodation and food service jobs, which are an indication of hotel and restaurant activity, are located all over Manhattan. It is not surprising to see that they are influential almost all day from pick-up and drop-off trip generation coefficients, but it has relatively higher influence at breakfast time (7 a.m. – 9a.m.), lunch time (1 p.m.), and dinner time (5 p.m. – 11:00 p.m.). These results provide us with a thorough understanding of the relationship between taxi demand and people's activities in Manhattan. If combined with other information, such as population, income, and TAT, the models provide predictions of taxi demand across time and space.

CONCLUSION

This study utilizes a large database of taxi trips with origins and destinations in NYC tracked by GPS and vast information on demographics and socioeconomics to build trip generation models at different times of day. A novel method was developed to calculate the minimum TAT using transit LOS and K-NN algorithm, and a procedure was implemented to select important explanatory variables using multiple linear regressions.

The TAT is mapped throughout 2,167 census tracts in NYC to compare with taxi demand, which clearly indicates the relationships between subway accessibility and taxi us. Taxi trips are more numerous in places where transit is more accessible. This relationship is confirmed in the multiple linear regression results. However, it is not possible to conclude with these methods whether the relationship between taxis and transit is competitive or complementary. Six major factors including population, education, age, income, TAT, and total jobs are shown to be influential in taxi trip generation modeling in NYC. Income and total jobs are the most influential factors because they are related to where people live and work.

The time of day modeling of taxi trips in Manhattan is used to identify several important factors from the job type category including the number of jobs at accommodation and food services and the number of jobs in retail. This information provides insights about where and when people start their activities and where and when they go home. The method presented in this paper creates a new way of interpreting trip generation modeling results. This approach to looking at trip making by time of day, and by pick-ups or drop-offs is helpful in understanding how the relationships between taxi demand and those influential factors vary temporally and spatially.

ACKNOWLEDGMENTS AND DISCLAIMER

This material is based upon work supported by the U.S. Department of Transportation's University Transportation Centers Program under Grant Number DTRT12-G-UTC21.

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

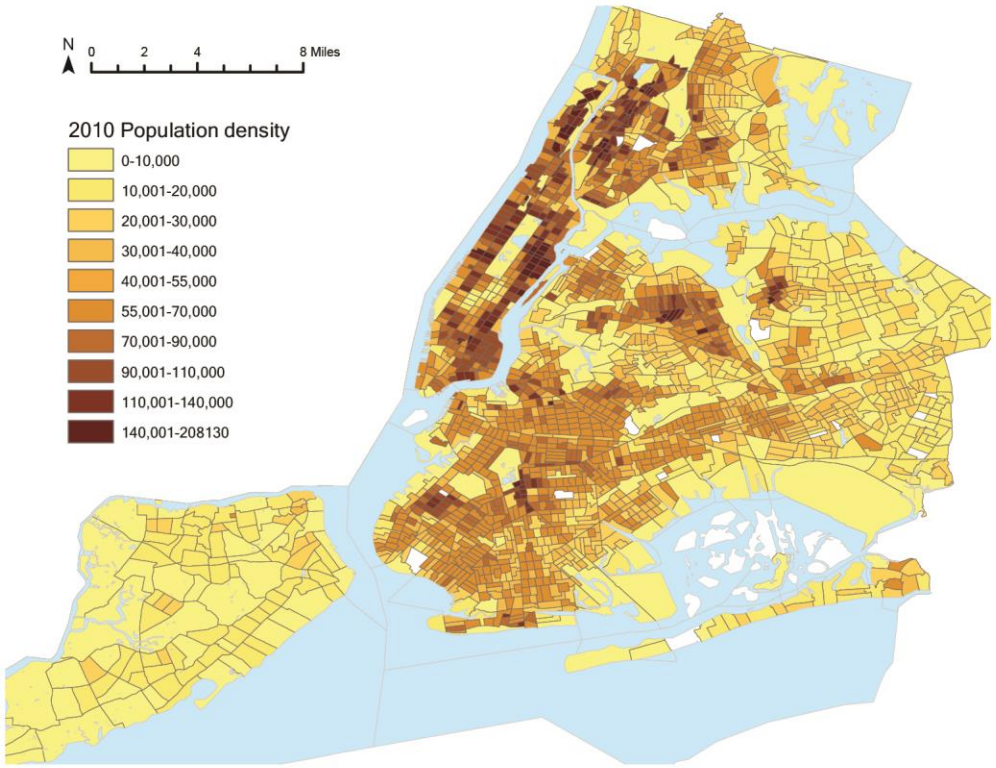
REFERENCES

1. Schaller Consulting, The New York City Taxicab Fact Book, 2006. Available on-line at http://www.schallerconsult.com/taxi/taxifb.pdf. Accessed on 11/1/2013.
2. O'Neill, W. A., and E. Brown. Long-Distance Trip Generation Modeling Using ATS. In *Transportation Research E-circular Number E-C026. Journal of Transportation Research Board*. Washington D.C., 2001.
3. Ben-Edigbe J., and R. Rahman. Multivariate School Travel Demand Regression Based on Trip Attraction. *World Academy of Science, Engineering and Technology*, Vol. 42, 2010, pp. 1169-1173.
4. Racca, D., and E. C. Ratledge. Project Report for Factors That Affect and/or Can Alter Mode Choice. Prepared for Delaware Transportation Institute and The State of Delaware Department of Transportation, 2004. Available on-line at http://udspace.udel.edu/handle/19716/1101. Accessed on June 12th 2013.
5. Kumar, A., and D. Levinson. Specifying, Estimating, and Validating a New Trip Generation Model: A Case Study of Montgomery County, Maryland. In *Transportation Research Record: Journal of Transportation Research Board, No. 1413*, 1992, pp. 107-113.
6. Schaller, B. A Regression Model of the Number of Taxicabs in U.S. Cities. *Journal of Public Transportation*, Vol. 8, 2005, pp. 63-78.
7. *Trip Generation Manual*, 9th Ed. Institute of Transportation Engineers, Washington, D.C., 2012.
8. Mousavi, A., J. Bunker, and B. Lee. A New Approach for Trip Generation Estimation for Traffic Impact Assessments. *25th ARRB Conference – Shaping the future: Linking policy research and outcomes*, Perth, Australia, 2012.
9. Yazici, M. A., C. Kamga, and K. Mouskos. Analysis of Travel Time Reliability in New York City Based on Day-of-Week Time-of-Day Periods. In *Transportation Research Record: Journal of Transportation Research Board, No. 2308*, 2012, pp. 83-95.
10. Corpuz, G., Public Transport or Private Vehicle: Factors that Impact on Mode Choice. In *Sydney, N.S.W. Transport Data Centre*, 30th Australasian Transportation Research Forum, 2007.
11. Chang, Y. C., Factors Affecting Airport Access Mode Choice for elderly air passengers. *Transportation Research Part E*, 2013 In press.
12. Schwanen, T., and P. L. Mokhtarian. What Affects Commute Mode Choice: Neighborhood Physical Structure or Preferences toward Neighborhoods? *Journal of Transport Geography*, Vol. 13, 2005, pp. 83-99.
13. Ewing, R., W. Schroeer, and W. Greene. School Location and Student Travel. In *Transportation Reserach Record*, Vol. 1895, 2004, pp. 55-63.
14. Schmöcker, J. D., M. A. Quddus, R. B. Noland, and M. G. H. Bell. Estimating Trip Generation of Elderly and Disabled people. In *Transportation Research Record: Journal of Transportation Research Board, No. 1924*, 2005, pp. 9-18.
15. Browning, R., E. Baker, J. Herron, and R. Kram. Effects of obesity and sex on the energetic cost and preferred speed of walking. *Journal of Applied Physiology*, Vol. 100, 2005, pp. 390-398.
16. Loos, N. *Value Creation in Leveraged Buyouts: Analysis of Factors Driving Private Equity Investment Performance*. German University Publishers (DUV), 2006.

17. Chen, X., P. B. Ender, M. Mitchell, and C. Wells. *Stata Web Books Regression with Stata: Chapter 2 - Regression Diagnostics*. UCLA Institute for Digital Research and Education, available online at http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm. Accessed on 11/1/2013.
18. Yan, X., and X. G. Su. *Linear Regression Analysis: Theory and Computing*. World Scientific Publishing Company, 1st edition, 2009.
19. Fox, J. *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications, Inc, 2nd edition, 2008.
20. Davies, A. Exhaustive Regression an Exploration of Regression-Based Data Mining Techniques Using Super Computation. *Research Program on Forecasting*, The George Washington University, RPF Working Paper No. 2008-008. Available online at http://www.gwu.edu/~forcpgm/2008-008.pdf Accessed on 7/27/2013.
21. McLeod, A., and C. Xu. Help bestglm: Best Subset GLM, available online at http://cran.r-project.org/web/packages/bestglm/vignettes/bestglm.pdf. Accessed on 7/26/2013.
22. Elliott, G., Gargano, A., Timmermann, A. Complete Subset Regressions, 2012. Available online at http://rady.ucsd.edu/docs/faculty/timmerman/subset-regression-April-25-2012.pdf. Accessed on 7/29/2013
23. Mallows, C. L., Some Comments on C_P. *Technometrics*, Vol. 15**,** 1975, pp. 661–675.
24. Funderburg, R. G., H. Nixon, M. G. Boarnet, and G. Ferguson. New highways and land use change: Results from a quasi-experimental design, *Transportation Research Part A*, Vol. 44, 2010, pp. 76-98.
25. Chatman, D. G. Does TOD Need the T? *Journal of the American Planning Association*, Vol. 79, 2013, pp. 17-31.

List of Figures

List of Tables



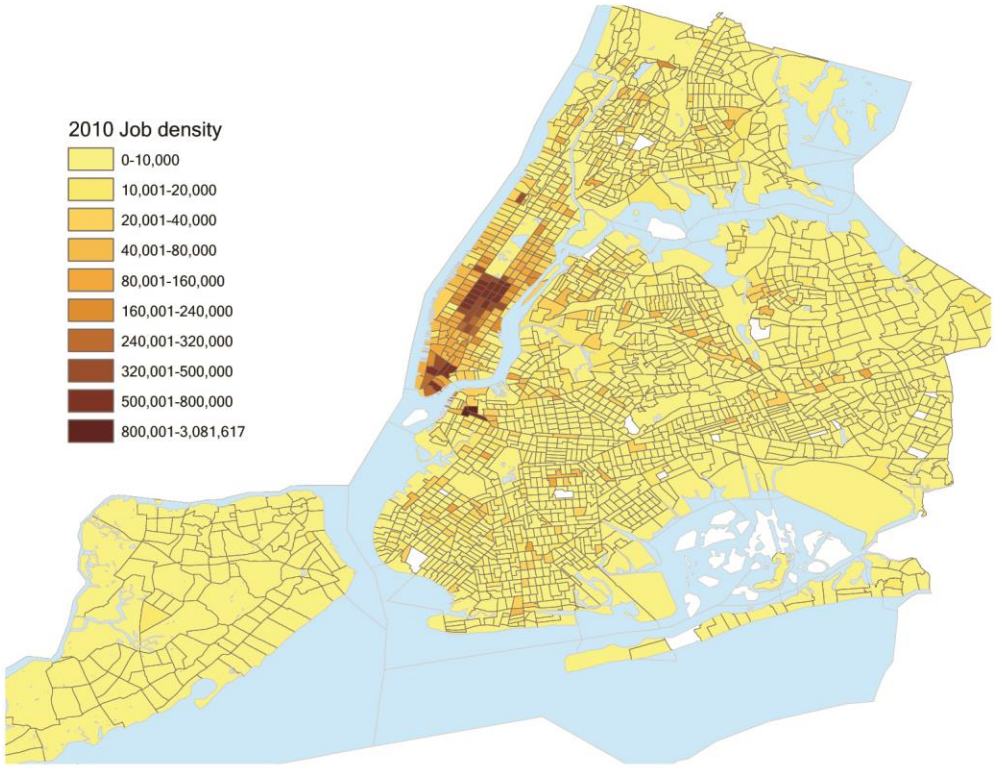


FIGURE 1 2010 population density and job density (per square mile)

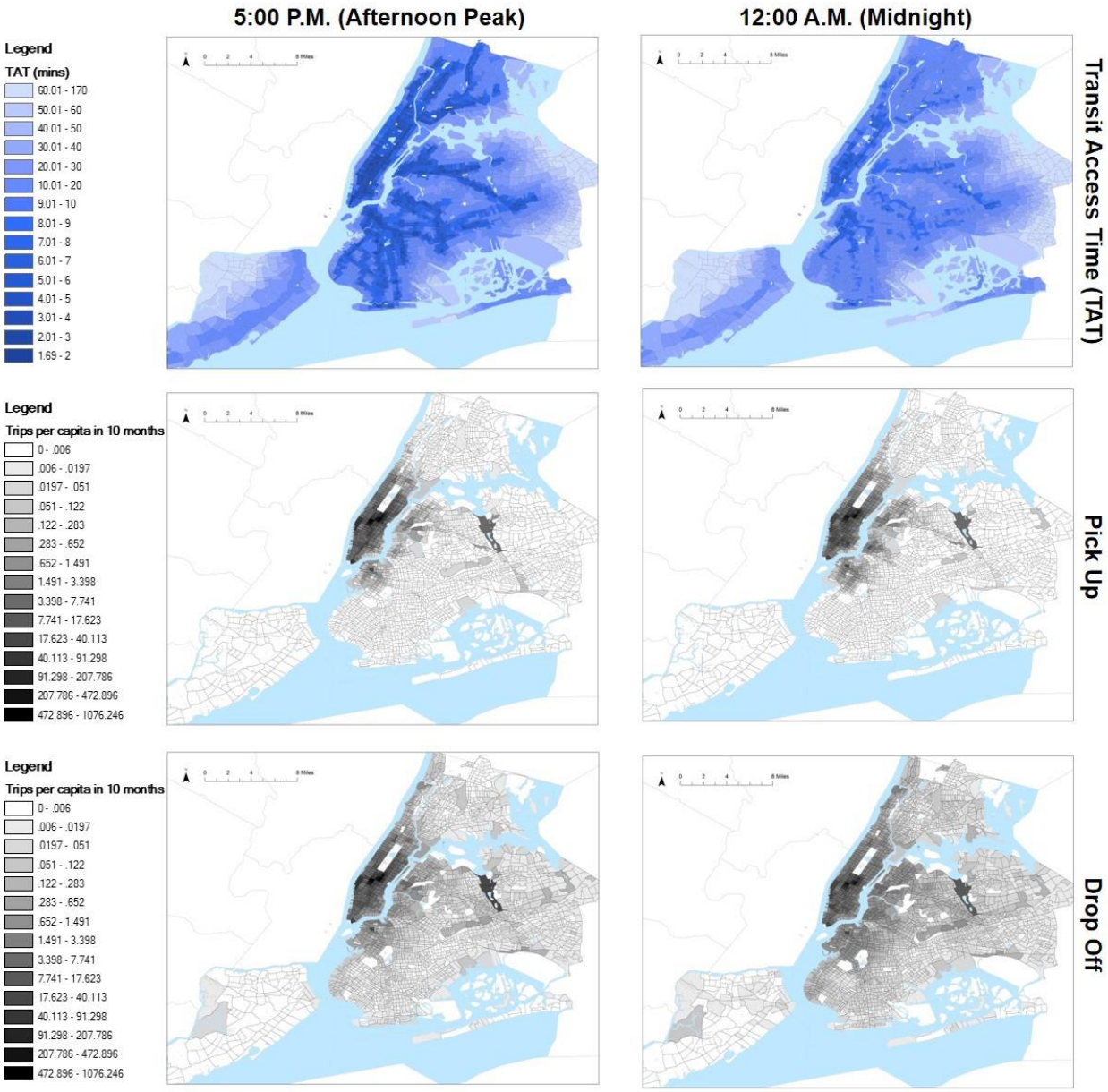


FIGURE 2 TAT and pick-up and drop-off taxi demand per capita at 5 p.m. and 12 a.m.

TABLE 1 List of explanatory variables in each model

Factor group	Factors or factor category	No. of variables	Initial model	Model with major factors	Manhattan model
TAT	TAT at specific hour*	1	√	√	√
Population	Total population (Pop)*	1	√	√	√
	Population by race	8	√	—	—
	Population by age	14	√	—	—
Age	Medium age (MedAge)*	1	√	√	√
Education	Percentage education higher than high school	1	√	—	—
	Percentage education higher than Bachelor (EduBac)*	1	√	√	√
Income	Median household income*	1	√	—	—
	Mean household income	1	√	—	—
	Median family income	1	√	—	—
	Mean family income*	1	√	—	—
	Per capita income (CapInc)*	1	√	√	√
Employment	Total jobs (TotJob)*	1	√	√	—
	Jobs by age	3	√	—	—
	Jobs by earnings	3	√	—	—
	Jobs by types*	20	√	—	√
	Jobs by race	6	√	—	—
	Jobs by ethnicity	2	√	—	—
	Jobs by education attainment	4	√	—	—
	Jobs by sex*	2	√	—	—
Total No. of variables		70	70	6	25

*****influential factors or factor category identified from step-wise selection (p-value < 0.05 or statistically significant at 95% level).

√ factor included in the model

— factor omitted from the model

TABLE 2 The coefficients of models for pick-ups and drop-offs in NYC

		Model fit statistics				Coefficients of explanatory variables					
	Hour	R2	AdjR2	Cp	BIC	Pop	MedAge	EduBac	CapInc	TAT	TotJob
Pick-ups	12 a.m.	0.47	0.46	5.87	-1248.26	0.48	-198.18	—	0.26	-36.38	0.32
	1 a.m.	0.38	0.38	4.00	-943.85	0.38	-142.05	—	0.19	-30.63	0.21
	2 a.m.	0.30	0.30	4.90	-704.99	0.31	-103.55	—	0.14	-23.54	0.13
	3 a.m.	0.26	0.26	5.67	-577.45	0.23	-72.69	—	0.10	-17.65	0.09
	4 a.m.	0.32	0.32	5.74	-753.91	0.17	-52.83	—	0.07	-11.97	0.07
	5 a.m.	0.52	0.52	5.14	-1464.34	0.16	-49.18	—	0.06	-7.56	0.06
	6 a.m.	0.48	0.48	6.00	-1303.97	0.35	-112.34	-25.36	0.15	-13.72	0.14
	7 a.m.	0.56	0.56	6.00	-1621.84	0.64	-210.90	-64.23	0.31	-23.79	0.24
	8 a.m.	0.61	0.61	6.00	-1872.85	0.70	-255.40	-99.03	0.41	-32.77	0.36
	9 a.m.	0.61	0.61	6.00	-1886.77	0.62	-249.00	-106.00	0.42	-37.68	0.43
	10 a.m.	0.62	0.62	6.00	-1959.95	0.57	-228.94	-103.30	0.39	-37.49	0.43
	11 a.m.	0.63	0.63	6.00	-1990.18	0.47	-216.05	-114.07	0.40	-39.79	0.49
	12 p.m.	0.63	0.63	6.00	-2002.51	0.44	-221.88	-125.07	0.43	-43.76	0.54
	1 p.m.	0.63	0.63	6.00	-2019.18	0.41	-216.54	-125.18	0.43	-44.67	0.53
	2 p.m.	0.63	0.63	6.00	-2013.17	0.40	-219.94	-132.22	0.44	-46.60	0.55
	3 p.m.	0.64	0.64	6.00	-2048.37	0.41	-213.26	-121.48	0.42	-43.62	0.49
	4 p.m.	0.64	0.64	6.00	-2059.71	0.39	-190.62	-101.24	0.36	-37.57	0.42
	5 p.m.	0.65	0.64	6.00	-2082.61	0.49	-237.27	-123.22	0.44	-44.38	0.50
	6 p.m.	0.64	0.64	6.00	-2034.77	0.57	-285.92	-155.81	0.54	-54.72	0.63
	7 p.m.	0.63	0.63	6.00	-1981.57	0.60	-300.80	-154.44	0.56	-56.44	0.67
	8 p.m.	0.62	0.61	6.00	-1915.26	0.55	-277.92	-129.10	0.50	-52.44	0.63
	9 p.m.	0.59	0.59	6.00	-1790.40	0.56	-266.61	-111.59	0.47	-52.15	0.60
	10 p.m.	0.56	0.56	6.00	-1642.60	0.56	-257.86	-92.14	0.44	-50.00	0.56
	11 p.m.	0.53	0.53	6.00	-1504.76	0.55	-236.71	-54.38	0.37	-44.68	0.45
Drop-offs	12 a.m.	0.60	0.59	6.00	-1809.31	0.67	-190.57	19.37*	0.22	-36.05	0.22
	1 a.m.	0.60	0.59	6.00	-1812.34	0.52	-136.56	25.23	0.14	-27.86	0.16
	2 a.m.	0.59	0.59	6.00	-1796.86	0.41	-100.30	24.64	0.10	-20.09	0.12
	3 a.m.	0.61	0.61	6.00	-1900.70	0.30	-68.96	18.14	0.06	-14.69	0.09
	4 a.m.	0.59	0.59	6.00	-1782.06	0.18	-40.04	10.99	0.04	-10.13	0.07
	5 a.m.	0.45	0.45	6.00	-1169.35	0.06	-22.93	-7.29	0.04	-7.71	0.11
	6 a.m.	0.43	0.43	4.02	-1120.11	—	-49.12	-54.92	0.14	-16.61	0.39
	7 a.m.	0.47	0.47	4.16	-1262.46	—	-95.27	-109.00	0.27	-32.15	0.71
	8 a.m.	0.53	0.53	4.00	-1528.36	—	-132.37	-129.14	0.36	-41.17	0.83
	9 a.m.	0.57	0.57	4.24	-1706.02	—	-140.65	-131.07	0.37	-44.19	0.76
	10 a.m.	0.60	0.60	6.00	-1840.99	0.17	-156.26	-114.32	0.36	-42.18	0.60
	11 a.m.	0.61	0.61	6.00	-1876.01	0.23	-168.98	-117.39	0.37	-43.79	0.56
	12 p.m.	0.62	0.62	6.00	-1952.91	0.31	-190.60	-122.02	0.40	-46.73	0.56
	1 p.m.	0.62	0.62	6.00	-1957.41	0.33	-192.48	-116.00	0.40	-45.16	0.55
	2 p.m.	0.62	0.62	6.00	-1923.38	0.39	-204.54	-116.39	0.41	-45.51	0.54
	3 p.m.	0.62	0.62	6.00	-1943.27	0.44	-207.17	-109.44	0.39	-43.12	0.47
	4 p.m.	0.62	0.62	6.00	-1958.00	0.45	-192.80	-90.60	0.35	-37.40	0.38
	5 p.m.	0.64	0.64	6.00	-2052.05	0.65	-251.19	-96.90	0.42	-42.86	0.42
	6 p.m.	0.65	0.65	6.00	-2122.94	0.86	-317.45	-106.75	0.50	-51.41	0.44
	7 p.m.	0.63	0.63	6.00	-2019.56	0.95	-338.33	-92.14	0.51	-52.89	0.43
	8 p.m.	0.64	0.64	6.00	-2047.76	0.98	-327.00	-56.46	0.45	-49.35	0.34
	9 p.m.	0.66	0.66	6.00	-2163.16	0.97	-312.84	-42.50	0.42	-48.29	0.33
	10 p.m.	0.66	0.66	6.00	-2182.20	0.95	-297.63	-26.59*	0.38	-45.89	0.33
	11 p.m.	0.63	0.63	4.00	-2026.33	0.84	-254.24	—	0.31	-42.39	0.29

*indicates non-significance of the coefficient, p-value>0.05

— factor omitted from the model

TABLE 3 The coefficients of models for drop-off trips in Manhattan

Drop-offs	Model fit statistics				Coefficients of explanatory variables												
Hour	R2	AdjR2	Cp	BIC	Pop	MedAge	EduBac	CapInc	JobCon	JobRet	JobTrW	JobFin	JobRea	JobPro	JobHea	JobEnt	JobFod
12 a.m.	0.81	0.81	24.64	-384.88	1.19	-273.30	221.63	—	10.95	—	—	-0.86	—	—	—	—	10.66
1 a.m.	0.81	0.81	27.16	-380.93	0.98	-212.91	160.73	—	8.49	—	—	-0.56	—	—	—	—	7.07
2 a.m.	0.81	0.80	30.52	-376.12	0.79	-160.57	122.33	—	8.41	—	—	—	-4.91	—	—	—	5.56
3 a.m.	0.83	0.82	31.02	-401.68	0.58	-112.09	82.79	—	6.32	—	—	—	-2.97	—	—	—	3.59
4 a.m.	0.81	0.80	36.56	-379.22	0.33	-55.91	42.52	—	—	—	3.39	—	—	—	0.29	—	1.98
5 a.m.	0.73	0.72	5.36	-293.50	0.11	—	—	—	—	0.89	—	—	—	0.79	0.42	0.70	1.49
6 a.m.	0.78	0.77	2.85	-348.22	—	—	—	—	—	—	—	1.25	4.27	0.81	1.16	—	6.74
7 a.m.	0.85	0.84	8.74	-437.84	—	—	—	—	—	—	—	1.80	15.24	1.96	1.69	—	10.28
8 a.m.	0.89	0.89	24.82	-529.64	—	—	—	—	—	5.35	—	—	16.13	4.66	1.50	—	11.27
9 a.m.	0.92	0.92	35.50	-599.30	—	—	—	0.07	—	8.27	—	—	—	5.84	1.14	—	10.49
10 a.m.	0.92	0.91	48.88	-587.50	—	—	—	0.10	—	9.15	—	—	—	4.00	1.23	—	8.19
11 a.m.	0.91	0.91	54.87	-579.39	—	—	—	0.11	—	9.95	—	—	—	3.08	1.14	—	9.06
12 p.m.	0.92	0.92	58.89	-604.07	—	—	—	0.14	—	9.50	—	—	—	2.88	—	3.22	9.24
1 p.m.	0.91	0.91	51.34	-583.66	—	—	—	0.13	—	8.14	—	—	—	2.48	1.21	—	11.17
2 p.m.	0.89	0.89	55.47	-523.75	—	—	—	0.15	—	8.50	—	—	—	2.69	1.35	—	9.97
3 p.m.	0.87	0.87	62.04	-485.79	—	—	—	0.17	—	7.77	—	—	—	2.70	—	3.79	6.42
4 p.m.	0.86	0.86	62.77	-467.86	—	—	—	0.16	—	5.59	—	—	—	2.12	—	3.46	5.94
5 p.m.	0.87	0.87	68.76	-480.04	—	—	—	0.23	—	—	—	-1.18	—	2.93	—	4.23	10.35
6 p.m.	0.88	0.88	49.48	-510.84	1.47	-369.24	—	0.33	—	—	—	—	—	—	1.56	—	17.11
7 p.m.	0.88	0.88	34.37	-498.22	1.66	-367.88	—	0.33	—	—	—	-1.20	—	—	—	—	21.48
8 p.m.	0.85	0.85	37.19	-443.27	1.71	-476.52	252.02	0.17	—	—	—	—	—	—	—	—	15.09
9 p.m.	0.86	0.86	33.25	-467.69	1.78	-477.56	250.63	0.15	—	—	—	—	—	—	—	—	14.02
10 p.m.	0.87	0.87	33.91	-476.50	1.75	-457.24	258.87	0.12	—	—	—	—	—	—	—	—	13.62
11 p.m.	0.85	0.85	30.83	-442.12	1.51	-377.97	314.88	—	—	—	9.76	—	—	—	—	—	12.38

'JobCon':Number of jobs in Construction; 'JobRet' Number of jobs in Retail Trade; 'JobTrW' Number of jobs in Transportation and Warehousing; 'JobFin' Number of jobs in Finance and Insurance; 'JobRea' Number of jobs in Real Estate and Rental and Enterprises; 'JobPro' Number of jobs in Professional, Scientific, and Enterprises; 'JobHea' Number of jobs in Health Care and Social Assistance; 'JobEnt' Number of jobs in Arts, Entertainment, and Recreation; 'JobFod' Number of jobs in Accommodation and Food Services.

— Factor omitted from the model

TABLE 4 The coefficients of models for pick-up trips in Manhattan

	Model fit statistics				Coefficients of explanatory variables												
Hour	R2	AdjR2	Cp	BIC	Pop	MedAge	EduBac	CapInc	TAT	JobCon	JobRet	JobTrW	JobFin	JobRea	JobPro	JobHea	JobEnt	JobFod
12 a.m.	0.78	0.77	11.68	-339.20	—	—	230.85	—	-431.29	15.88	—	—	-1.01	-13.25	—	—	—	20.26
1 a.m.	0.69	0.68	11.71	-255.41	—	—	177.18	—	-285.36	13.91	—	—	-1.01	-14.24	—	—	—	16.47
2 a.m.	0.61	0.60	9.49	-202.36	—	—	97.03	—	—	13.48	—	—	-1.11	-14.34	—	—	-2.49	15.23
3 a.m.	0.57	0.56	7.09	-176.59	—	—	70.29	—	—	10.51	—	—	-0.88	-11.85	—	—	-2.42	11.85
4 a.m.	0.66	0.65	17.14	-231.99	—	—	48.85	—	—	7.25	—	—	-0.56	-7.30	—	—	-1.30	7.81
5 a.m.	0.76	0.75	25.55	-321.28	0.29	—	50.97	—	-229.78	—	—	3.41	—	—	—	0.31	—	2.26
6 a.m.	0.65	0.65	13.51	-231.69	0.82	-177.08	—	0.13	—	—	—	—	—	—	1.53	0.80	2.41	—
7 a.m.	0.74	0.73	15.21	-302.89	1.50	-348.46	—	0.25	—	—	—	—	—	—	2.76	1.27	3.87	—
8 a.m.	0.81	0.80	29.08	-378.73	1.52	-395.80	—	0.32	—	—	—	—	—	—	4.26	1.46	4.80	—
9 a.m.	0.83	0.82	43.97	-404.27	1.26	-360.67	—	0.29	—	—	—	—	—	—	3.07	1.66	—	8.30
10 a.m.	0.86	0.86	57.19	-454.41	1.01	-308.34	—	0.24	—	—	—	—	—	15.05	—	1.63	—	8.87
11 a.m.	0.89	0.88	61.13	-508.46	—	—	—	0.17	—	—	—	—	—	11.95	2.04	1.16	3.95	6.55
12 p.m.	0.90	0.90	67.32	-538.85	—	—	—	0.17	—	—	7.23	—	—	—	3.46	1.02	3.54	7.39
1 p.m.	0.91	0.91	66.95	-564.52	—	-141.02	—	0.23	—	—	9.17	—	—	—	2.61	1.26	—	9.15
2 p.m.	0.91	0.91	63.60	-580.34	—	—	—	0.16	—	—	8.98	—	—	—	3.00	0.96	3.68	8.35
3 p.m.	0.91	0.91	64.45	-567.34	—	-154.67	—	0.22	—	—	8.03	—	—	10.51	—	1.37	—	9.42
4 p.m.	0.91	0.90	64.95	-554.60	—	-93.66	—	0.20	—	—	6.89	—	—	—	1.87	—	3.21	6.96
5 p.m.	0.91	0.90	63.58	-555.30	0.98	-288.01	—	0.24	—	—	8.46	—	—	—	2.31	—	—	10.73
6 p.m.	0.92	0.91	58.17	-583.47	—	—	—	0.22	—	—	6.59	—	-1.11	—	4.10	—	4.94	13.84
7 p.m.	0.92	0.91	55.86	-584.17	—	—	—	0.22	—	—	5.94	—	-1.24	—	4.04	—	5.30	16.35
8 p.m.	0.91	0.90	54.69	-557.68	—	-250.61	335.81	—	—	—	7.85	—	-1.19	—	3.32	—	—	18.86
9 p.m.	0.90	0.89	35.05	-528.50	0.93	-382.70	310.79	—	—	—	7.11	—	—	—	1.72	—	—	19.76
10 p.m.	0.88	0.88	21.45	-495.67	0.84	-363.06	289.95	—	—	—	5.46	—	—	—	—	0.84	—	22.69
11 p.m.	0.84	0.84	19.30	-428.32	0.77	-332.23	285.40	—	—	—	4.16	—	-0.93	—	—	—	—	21.11