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

Ci Yang and Eric J. Gonzales

Identifying the factors that influence taxi demand is very important for understanding where and when people use taxis. A large set of GPS data from New York City taxis is used along with demographic, socioeconomic, and employment data to identify the factors that drive taxi demand. A technique was developed to measure and map transit accessibility on the basis of transit access time (TAT) to understand the relationship between taxi use and transit service. The taxi data were categorized by pickups and drop-offs at different times of day. A multiple linear regression model was estimated for each hour of the day to model pickups and another to model drop-offs. Six important explanatory variables that influence taxi trips were identified: population, education, age, income, TAT, and employment. The influence of these factors on taxi pickups and drop-offs changed at different times of the day. The number of jobs in each industry sector was an indication of the types of economic activities occurring at a location, and in some sectors the number of jobs were 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.

Taxis in New York City carry 172 million trips annually (11% of all travel) this fact makes the cabs an important transport mode in the city (*1*). All New York City 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. For effective planning and management of the taxi fleet, understanding 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 is necessary. A trip generation model that relates taxi demand to the 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 process 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, 4–6*):

- Level of service (LOS) of the mode,
- Accessibility of the mode,
- Demographics of the TAZ (e.g., population and race),
- Socioeconomics of the TAZ (e.g., income and education),
- Other characteristics of the TAZ (e.g., land area), and
- Land use in the TAZ.

Three methods are commonly used to model trip production: the rate method (*7*), cross classification (*2, 5*), and regression (*3, 8*). The rate method is used for traffic impact analysis on nonresidential 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 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 data set with sufficiently detailed information about travel and TAZ characteristics is necessary to model trip generation across a large geographic area by using regression.

In this study, taxi trip information from 10 months of complete GPS data is related with transit information in New York City. 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, but these require that detailed transit schedule information be cleaned and compiled before inclusion in the model (*2, 3*). Detailed taxi and transit data that have spatial and temporal components allow for investigation of how the factors that drive taxi use change at different times of the day.

This paper is organized as follows. First, the literature on factors that have been found to influence trip generation is reviewed, followed by a section on data that provides a description of the taxi GPS data and the explanatory variables in this study. Then

C. Yang, Department of Civil and Environmental Engineering, Rutgers University, 96 Frelinghuysen Road, Piscataway, NJ 08854. E. J. Gonzales, Department of Civil and Environmental Engineering, University of Massachusetts, Amherst, 130 Natural Resources Road, Amherst, MA 01003. Corresponding author: C. Yang, jessie_yang06@yahoo.com.

Transportation Research Record: Journal of the Transportation Research Board, No. 2429, Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 110–120.
DOI: 10.3141/2429-12

a novel technique to calculate transit accessibility and a multiple linear regression model to identify influential factors are described. The model results are presented following the methodology, and conclusions are discussed.

LITERATURE REVIEW

There are few studies using large taxi GPS data sets to model taxi trip generation. Schaller presents an analysis of the number of taxicabs in 118 U.S. cities using multiple linear regression models (*6*). 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. 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 (*8*).

Taxi demand may be closely related to transit accessibility in New York City because taxis and transit 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 that reason, factors related to transit and vehicle modal split could be included in models for taxi demand. Racca and Ratledge present a comprehensive list of possible factors that are used for mode choice modeling, including transit LOS, accessibility, land use, demographics, and trip characteristics (*4*). 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 of mode split versus mean age and time of day indicates that these variables affect the modes that people choose, and this means that they may also relate to taxi trip generation. Corpuz shows that socioeconomic characteristics and time of day have influenced people's choices between private vehicles and public transportation (*9*). 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 chosen during morning and late afternoon peaks because people want to avoid the time and the cost of driving in congestion (*9*).

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, 9, 10*). Trips to residential areas and nonresidential areas (*11*) and trips for business and nonbusiness purposes (*2*) are analyzed separately in most studies. A number of studies have been conducted concerning the generation of airport trips (*10*) and travel to schools (*3, 12*). Researchers have also studied trips generated by elderly people because their needs and behavior have some distinct differences from those of other population groups (*10, 13*).

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 of the characteristics of the places where taxi trips start and end to gain insights into 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 to develop models for taxi trip generation.

DATA

Trip generation models require comprehensive sets of data for explanatory variables to identify the most influential factors on taxi trip generation. The database of taxi trips has complete information on 147 million taxi trips made between February 1, 2010, and November 28, 2010, including temporal and spatial information acquired by GPS (taxi pickup and drop-off date, time, and location), fare (including tolls, tip, and total fare paid), and distance traveled. The taxi data for pickup and drop-off locations are aggregated by hour of the day in a manner similar to the way that taxis were used as traffic probes by time of day in Yazici et al. (*14*). The distribution of pickups (origins) and drop-offs (destinations) is 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

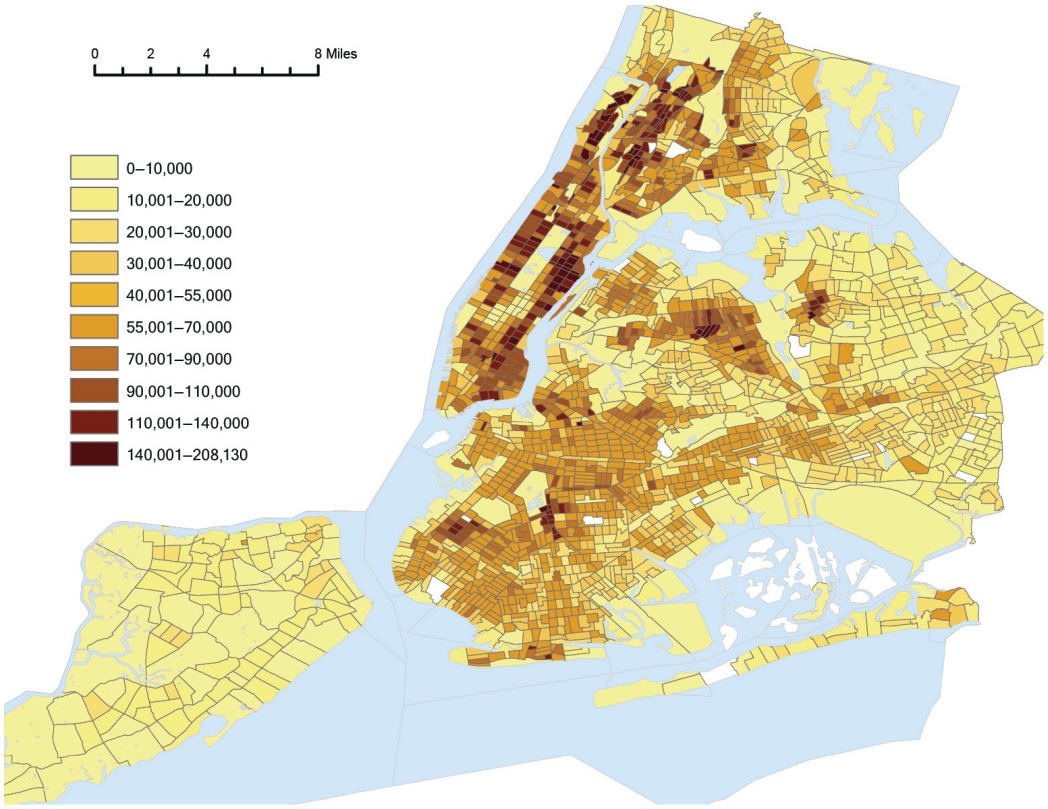
- Transit LOS based on New York City subway schedules available from Google transit feed data in the format of general transit feed specification;
- Demographics data for each census tract available from the U.S. Census 2010, including total population, population categorized by age, and population categorized by race;
- Socioeconomic data 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 available for New York City from 2010 workplace area characteristic data from the U.S. Bureau of the Census; and
- Geographic data including relevant shapefiles (i.e., rivers, roads, county, and 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 are calculated for all 2,167 census tracts in New York City. Figure 1 shows that the population density and employment density are concentrated in Manhattan. When the taxi demand information in Figure 2, is compared, pickup demand clearly appears concentrated in Manhattan, northern Brooklyn, and the west and north sides of Queens, whereas the drop-off demand is more spread over four of the boroughs: Manhattan, Brooklyn, Queens, and the Bronx.

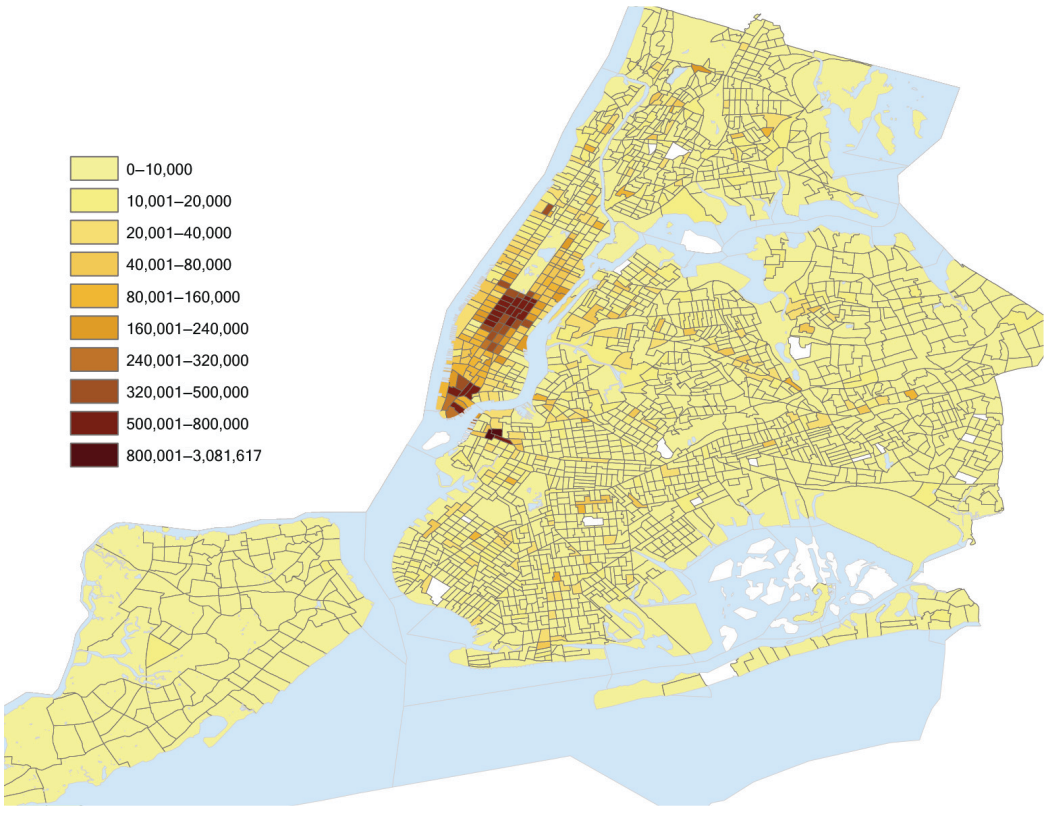
Some census tracts consisting of cemeteries, parks, or islands do not have employment associated with them, so the workplace area characteristic data cover 2,143 census tracts. Census tracts with variables that are lacking certain required information are excluded from the linear model analysis. Ultimately, 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 procedure



(a)



(b)

FIGURE 1 2010 New York City demographic densities (per square mile): (*a*) population and (*b*) jobs.

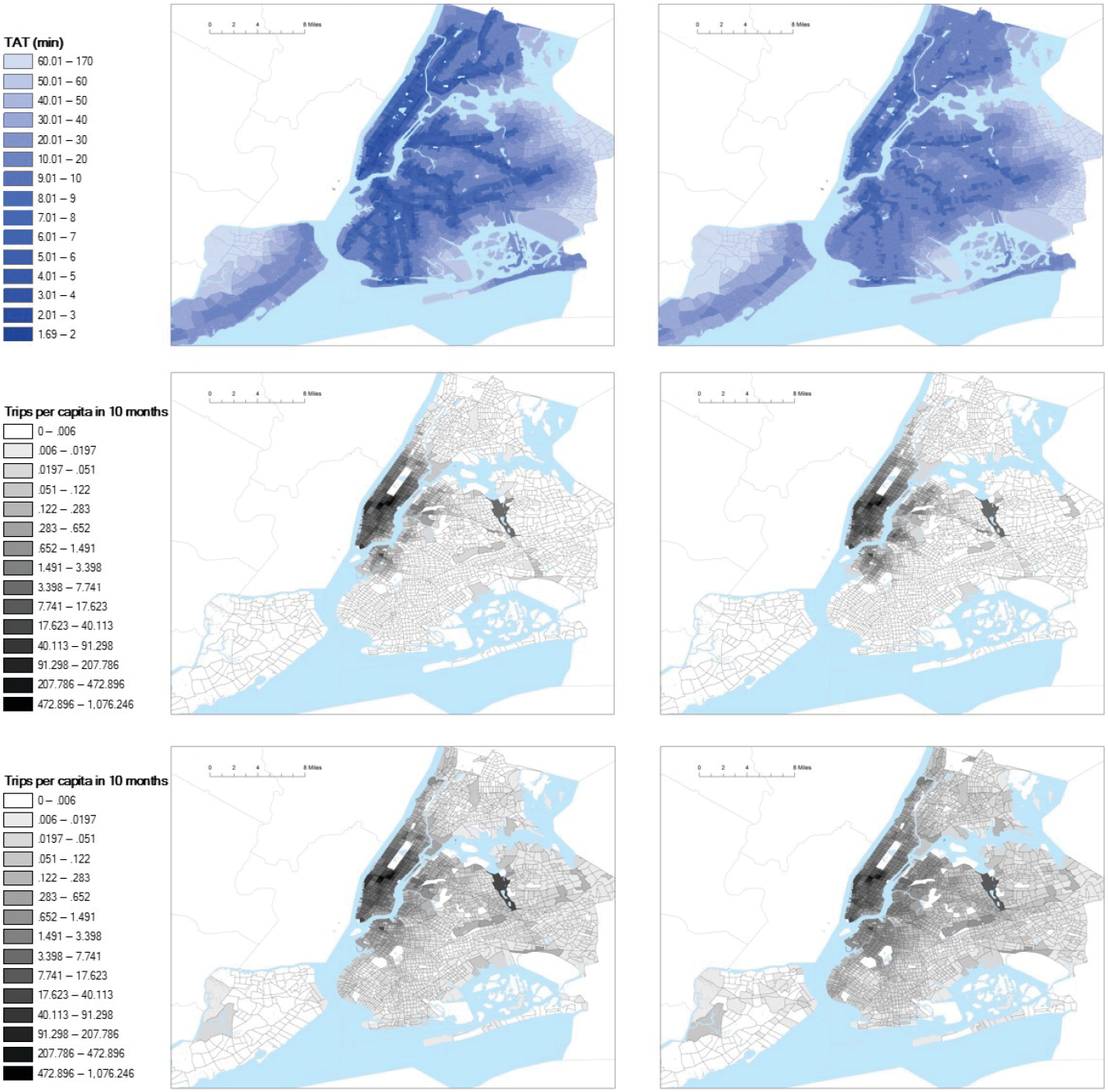


FIGURE 2 Transit access time (*top*) and pickup (*middle*) and drop-off (*bottom*) taxi demand per capita at (*a*) 5 p.m. and (*b*) midnight.

requires processing raw transit schedule information to determine how much time it takes a 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 or regression model for estimating taxi trip generation. The taxi data are cross classified by pickup and drop-off and aggregated by hour of the day. In 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 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 km/h), the TAT in minutes is as follows (*15*):

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

where

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

The minimum TAT is calculated at each location by the following steps. First, the transit schedule in general transit feed specification 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 m (820 ft) square, which is small enough that the walking time to cross each cell is less than 1 min. 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 New York City at each hour of the day, and that result is used to quantify transit accessibility in the city with spatial resolution of 250 m (820 ft) and temporal resolution of an hour.

Once the minimum TAT for each census tract is determined, it is not difficult to calculate the TAT by averaging the values across the cells included in the census tract. This method 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 by using only the subway data in this study because the complete general transit feed specification bus schedule data are incomplete (e.g., bus data for Queens are not available).

Visualizing TAT and Taxi Demand

Figure 2 shows the TAT for subways at midnight and 5:00 p.m. (afternoon) along with taxi pickups 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 according to the spatial coverage of the subway network. The transit accessibility is also generally greater at 5 p.m. than at midnight, because services operate more frequently during the peak hours than late at night. Figure 2 suggests that the pickups 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 hour of the day because the distribution of activities in New York City changes with time. There are also differences between the rates of taxi pickups per capita at 5 p.m. and at midnight. For example, there are more taxi pickups at Jamaica at 5 p.m. than at midnight, which could result from people getting off the subway at Jamaica and then taking a taxi to complete a trip home from work. In some areas of lower Manhattan there are more pickups 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 midnight 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 midnight, which is consistent with the fact that these are busy transit hubs used by commuters. Although the total amount of travel activity in the city is lower at midnight than at 5 p.m., many areas of the outer boroughs actually see a greater rate of drop-offs in the late night hours. This finding 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 pickups 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 provides a visualization of their relationship and helps provide intuition about why such a relationship exists. With the hourly data for TAT, taxi pickups, 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 to achieve that 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:

$$Y = \sum_{i=1}^{n} \beta_i X_i + \beta_0 + \varepsilon \quad (2)$$

where

Y = number of taxi trips generated in a TAZ (response variable),
X_i = n independent variables,
β_0 = intercept,
β_i = coefficient corresponding to X_i, and
ε = error representing the difference between modeled and observed number of taxi trips.

The response variable in the model is the number of pickups or drop-offs generated in each census tract by hour of the day from the 10-month taxi GPS data in New York City. The explanatory variables considered in the initial model are listed in Table 1. With least squares estimation (i.e., maximum likelihood estimate) coefficients are estimated for each explanatory variable by minimizing the mean squared error between the modeled Y and observed Y. The goal is to 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, stepwise selection and

TABLE 1 Explanatory Variables in Each Model

Factor Group	Factors or Factor Category	Number of Variables	Initial Model	Model with Major Factors	Manhattan Model
TAT	TAT at specific hour[a]	1	√	√	√
Population	Total population (Pop)[a]	1	√	√	√
	Population by race	8	√	—	—
	Population by age	14	√	—	—
Age	Medium age (MedAge)[a]	1	√	√	√
Education	Percentage education higher than high school	1	√	—	—
	Percentage education higher than bachelor (EduBac)[a]	1	√	√	√
Income	Median household income[a]	1	√	—	—
	Mean household income	1	√	—	—
	Median family income	1	√	—	—
	Mean family income[a]	1	√	—	—
	Per capita income (CapInc)[a]	1	√	√	√
Employment	Total jobs (TotJob)[a]	1	√	√	—
	Jobs by age	3	√	—	—
	Jobs by earnings	3	√	—	—
	Jobs by types[a]	20	√	—	√
	Jobs by race	6	√	—	—
	Jobs by ethnicity	2	√	—	—
	Jobs by education attainment	4	√	—	—
	Jobs by sex[a]	2	√	—	—
Total number of variables		70	70	6	25

NOTE: √ = factor included; — = factor omitted.
[a]Influential factors or factor category identified from stepwise selection (p-value < .05 or statistically significant at 95% level).

best subset regression are two methods for comparing model specifications to identify the best set of explanatory variables to include in the final model. Several procedures used to select important variables in this study are described below.

Check Correlation Coefficients

An analysis of the correlation coefficients of the response variable and all explanatory variables shows how closely each pair of variables varies 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 that the explanatory variable is important. Strong 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 as a result of multicollinearity (*16, 17*). Each indicator has a variance inflation factor 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 variance inflation factor of each factor is larger than 5, then multicollinearity is high (*16, 17*).

Stepwise Selection

Stepwise 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 Akaike information criterion (AIC) value and a smaller residual sum of squares. 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 models with a large number of explanatory variables are compared. The stepwise method involves ranking the importance of each factor by listing the AIC values that 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 (also called complete subset regression) is a method to select the best subset of predictors from among all possible combinations of predictors (2^k combinations if there are k predictors in the initial model) (*20–22*). 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.
- The Bayesian information criterion, 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 compared with AIC is that the penalty term for the number of variables included in the model is larger in the Bayesian information criterion than in AIC (*18, 19*). For both metrics, lower values are an indication of a better model.

- Mallows's C_p assesses overfitting of the model, and a desirable model has a 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 pickups and drop-offs at different times of day on the basis of the same set of explanatory variables because trips at different times of day could be associated with different explanatory factors.

An R^2 greater than .8 in most trip generation studies is difficult to achieve because there are many things affecting the response variable, and the simplest possible model is sought 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 .5 and sometimes less than .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 one 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 on the basis not only of R^2 but also of other criteria used to select an appropriate model. 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 < .05 is significant at the 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 by using stepwise selection based on AIC values and residual sum of squares: TAT, total population, median age, three types of income, total jobs, jobs by type, and jobs by sex, which are listed in Table 1.

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. Because of the 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 is 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 from among two or more correlated factors is included.

Models with and without the intercept are estimated for pickups and drop-offs for each hour of the day in New York City. 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 six major variables for each time of the day, are presented in Table 2. All coefficients are significant (p-value < .05) unless labeled otherwise for Tables 2, 3, and 4.

The interpretation of the trip generation results for both pickups and drop-offs is useful for transportation planning and regulation of taxi services. The magnitude and sign of the coefficient for each explanatory variable indicate 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 pickups at 12 a.m. in New York City (Table 2), an indication 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 during a 10-month period. Similarly, there is an average decrease of 36 taxi trips at the same hour during a 10-month period as TAT increases by 1 min; this relationship provides insight into 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 information gives some idea of where and when more taxi use would be expected than actually occurs, based on citywide trends, so the information can be useful for planning taxi stand locations or providing incentives for cab drivers to operate during certain times of the 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.

Results show that population, education, income, and total jobs positively influence both taxi pickups and drop-offs in New York City. This finding 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 trip making by taxis. That finding shows that younger people are more likely to take taxis. Results also show that taxi demand is high where transit is more accessible (TAT is small). The available data do not clearly indicate 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 whether the large number of taxi trips is 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.; this result indicates that extra taxi demand during this period in New York City 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 many of the trips are possibly being made to or from transit facilities and 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 pickups and drop-offs in Manhattan, as listed in Tables 3 and 4. TAT loses its influence for drop-off trips in Manhattan compared with instances in which total jobs was used. Factors related to job types seem to play key roles in generating taxi trips in Manhattan; some influential industry sectors

TABLE 2 Coefficients of Models for Pickups and Drop-Offs in New York City

	Model-Fit Statistics				Coefficients of Explanatory Variables					
Hour	R^2	AdjR^2	C_p	BIC	Pop	MedAge	EduBac	CapInc	TAT	TotJob
Pickups										
midnight	.47	.46	5.87	−1,248.26	0.48	−198.18	—	0.26	−36.38	0.32
1 a.m.	.38	.38	4.00	−943.85	0.38	−142.05	—	0.19	−30.63	0.21
2 a.m.	.30	.30	4.90	−704.99	0.31	−103.55	—	0.14	−23.54	0.13
3 a.m.	.26	.26	5.67	−577.45	0.23	−72.69	—	0.10	−17.65	0.09
4 a.m.	.32	.32	5.74	−753.91	0.17	−52.83	—	0.07	−11.97	0.07
5 a.m.	.52	.52	5.14	−1,464.34	0.16	−49.18	—	0.06	−7.56	0.06
6 a.m.	.48	.48	6.00	−1,303.97	0.35	−112.34	−25.36	0.15	−13.72	0.14
7 a.m.	.56	.56	6.00	−1,621.84	0.64	−210.90	−64.23	0.31	−23.79	0.24
8 a.m.	.61	.61	6.00	−1,872.85	0.70	−255.40	−99.03	0.41	−32.77	0.36
9 a.m.	.61	.61	6.00	−1,886.77	0.62	−249.00	−106.00	0.42	−37.68	0.43
10 a.m.	.62	.62	6.00	−1,959.95	0.57	−228.94	−103.30	0.39	−37.49	0.43
11 a.m.	.63	.63	6.00	−1,990.18	0.47	−216.05	−114.07	0.40	−39.79	0.49
noon	.63	.63	6.00	−2,002.51	0.44	−221.88	−125.07	0.43	−43.76	0.54
1 p.m.	.63	.63	6.00	−2,019.18	0.41	−216.54	−125.18	0.43	−44.67	0.53
2 p.m.	.63	.63	6.00	−2,013.17	0.40	−219.94	−132.22	0.44	−46.60	0.55
3 p.m.	.64	.64	6.00	−2,048.37	0.41	−213.26	−121.48	0.42	−43.62	0.49
4 p.m.	.64	.64	6.00	−2,059.71	0.39	−190.62	−101.24	0.36	−37.57	0.42
5 p.m.	.65	.64	6.00	−2,082.61	0.49	−237.27	−123.22	0.44	−44.38	0.50
6 p.m.	.64	.64	6.00	−2,034.77	0.57	−285.92	−155.81	0.54	−54.72	0.63
7 p.m.	.63	.63	6.00	−1,981.57	0.60	−300.80	−154.44	0.56	−56.44	0.67
8 p.m.	.62	.61	6.00	−1,915.26	0.55	−277.92	−129.10	0.50	−52.44	0.63
9 p.m.	.59	.59	6.00	−1,790.40	0.56	−266.61	−111.59	0.47	−52.15	0.60
10 p.m.	.56	.56	6.00	−1,642.60	0.56	−257.86	−92.14	0.44	−50.00	0.56
11 p.m.	.53	.53	6.00	−1,504.76	0.55	−236.71	−54.38	0.37	−44.68	0.45
Drop-offs										
midnight	.60	.59	6.00	−1,809.31	0.67	−190.57	19.37[a]	0.22	−36.05	0.22
1 a.m.	.60	.59	6.00	−1,812.34	0.52	−136.56	25.23	0.14	−27.86	0.16
2 a.m.	.59	.59	6.00	−1,796.86	0.41	−100.30	24.64	0.10	−20.09	0.12
3 a.m.	.61	.61	6.00	−1,900.70	0.30	−68.96	18.14	0.06	−14.69	0.09
4 a.m.	.59	.59	6.00	−1,782.06	0.18	−40.04	10.99	0.04	−10.13	0.07
5 a.m.	.45	.45	6.00	−1,169.35	0.06	−22.93	−7.29	0.04	−7.71	0.11
6 a.m.	.43	.43	4.02	−1,120.11	—	−49.12	−54.92	0.14	−16.61	0.39
7 a.m.	.47	.47	4.16	−1,262.46	—	−95.27	−109.00	0.27	−32.15	0.71
8 a.m.	.53	.53	4.00	−1,528.36	—	−132.37	−129.14	0.36	−41.17	0.83
9 a.m.	.57	.57	4.24	−1,706.02	—	−140.65	−131.07	0.37	−44.19	0.76
10 a.m.	.60	.60	6.00	−1,840.99	0.17	−156.26	−114.32	0.36	−42.18	0.60
11 a.m.	.61	.61	6.00	−1,876.01	0.23	−168.98	−117.39	0.37	−43.79	0.56
noon	.62	.62	6.00	−1,952.91	0.31	−190.60	−122.02	0.40	−46.73	0.56
1 p.m.	.62	.62	6.00	−1,957.41	0.33	−192.48	−116.00	0.40	−45.16	0.55
2 p.m.	.62	.62	6.00	−1,923.38	0.39	−204.54	−116.39	0.41	−45.51	0.54
3 p.m.	.62	.62	6.00	−1,943.27	0.44	−207.17	−109.44	0.39	−43.12	0.47
4 p.m.	.62	.62	6.00	−1,958.00	0.45	−192.80	−90.60	0.35	−37.40	0.38
5 p.m.	.64	.64	6.00	−2,052.05	0.65	−251.19	−96.90	0.42	−42.86	0.42
6 p.m.	.65	.65	6.00	−2,122.94	0.86	−317.45	−106.75	0.50	−51.41	0.44
7 p.m.	.63	.63	6.00	−2,019.56	0.95	−338.33	−92.14	0.51	−52.89	0.43
8 p.m.	.64	.64	6.00	−2,047.76	0.98	−327.00	−56.46	0.45	−49.35	0.34
9 p.m.	.66	.66	6.00	−2,163.16	0.97	−312.84	−42.50	0.42	−48.29	0.33
10 p.m.	.66	.66	6.00	−2,182.20	0.95	−297.63	−26.59[a]	0.38	−45.89	0.33
11 p.m.	.63	.63	4.00	−2,026.33	0.84	−254.24	—	0.31	−42.39	0.29

NOTE: BIC = Bayesian information criterion; — = factor omitted from the model.
[a]Indicates nonsignificance of the coefficient, p-value > .05.

include jobs in retail, accommodation and food service, and health care (see Tables 3 and 4).

From 11 p.m. to 8 a.m., it appears that the drop-off taxi demand is not significantly related to income whereas from 9 a.m. to 10 p.m. it is. The indication is 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 pickup 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 noon 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. These

TABLE 3 Coefficients of Models for Drop-Off Trips in Manhattan

Drop-Off Hour	Model-Fit Statistics				Coefficients of Explanatory Variables												
	R^2	AdjR^2	C_p	BIC	Pop	MedAge	EduBac	CapInc	JobCon	JobRet	JobTrW	JobFin	JobRea	JobPro	JobHea	JobEnt	JobFod
midnight	.81	.81	24.64	−384.88	1.19	−273.30	221.63	—	10.95	—	—	−0.86	—	—	—	—	10.66
1 a.m.	.81	.81	27.16	−380.93	0.98	−212.91	160.73	—	8.49	—	—	−0.56	—	—	—	—	7.07
2 a.m.	.81	.80	30.52	−376.12	0.79	−160.57	122.33	—	8.41	—	—	—	−4.91	—	—	—	5.56
3 a.m.	.83	.82	31.02	−401.68	0.58	−112.09	82.79	—	6.32	—	—	—	−2.97	—	—	—	3.59
4 a.m.	.81	.80	36.56	−379.22	0.33	−55.91	42.52	—	—	—	3.39	—	—	—	0.29	—	1.98
5 a.m.	.73	.72	5.36	−293.50	0.11	—	—	—	—	0.89	—	—	—	0.79	0.42	0.70	1.49
6 a.m.	.78	.77	2.85	−348.22	—	—	—	—	—	—	—	1.25	4.27	0.81	1.16	—	6.74
7 a.m.	.85	.84	8.74	−437.84	—	—	—	—	—	—	—	1.80	15.24	1.96	1.69	—	10.28
8 a.m.	.89	.89	24.82	−529.64	—	—	—	—	—	5.35	—	—	16.13	4.66	1.50	—	11.27
9 a.m.	.92	.92	35.50	−599.30	—	—	—	0.07	—	8.27	—	—	—	5.84	1.14	—	10.49
10 a.m.	.92	.91	48.88	−587.50	—	—	—	0.10	—	9.15	—	—	—	4.00	1.23	—	8.19
11 a.m.	.91	.91	54.87	−579.39	—	—	—	0.11	—	9.95	—	—	—	3.08	1.14	—	9.06
noon	.92	.92	58.89	−604.07	—	—	—	0.14	—	9.50	—	—	—	2.88	—	3.22	9.24
1 p.m.	.91	.91	51.34	−583.66	—	—	—	0.13	—	8.14	—	—	—	2.48	1.21	—	11.17
2 p.m.	.89	.89	55.47	−523.75	—	—	—	0.15	—	8.50	—	—	—	2.69	1.35	—	9.97
3 p.m.	.87	.87	62.04	−485.79	—	—	—	0.17	—	7.77	—	—	—	2.70	—	3.79	6.42
4 p.m.	.86	.86	62.77	−467.86	—	—	—	0.16	—	5.59	—	—	—	2.12	—	3.46	5.94
5 p.m.	.87	.87	68.76	−480.04	—	—	—	0.23	—	—	—	−1.18	—	2.93	—	4.23	10.35
6 p.m.	.88	.88	49.48	−510.84	1.47	−369.24	—	0.33	—	—	—	—	—	—	1.56	—	17.11
7 p.m.	.88	.88	34.37	−498.22	1.66	−367.88	—	0.33	—	—	—	−1.20	—	—	—	—	21.48
8 p.m.	.85	.85	37.19	−443.27	1.71	−476.52	252.02	0.17	—	—	—	—	—	—	—	—	15.09
9 p.m.	.86	.86	33.25	−467.69	1.78	−477.56	250.63	0.15	—	—	—	—	—	—	—	—	14.02
10 p.m.	.87	.87	33.91	−476.50	1.75	−457.24	258.87	0.12	—	—	—	—	—	—	—	—	13.62
11 p.m.	.85	.85	30.83	−442.12	1.51	−377.97	314.88	—	—	—	9.76	—	—	—	—	—	12.38

NOTE: 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 Coefficients of Models for Pickup Trips in Manhattan

Pickup Hour	Model-Fit Statistics				Coefficients of Explanatory Variables													
	R^2	AdjR^2	C_p	BIC	Pop	MedAge	EduBac	CapInc	TAT	JobCon	JobRet	JobTrW	JobFin	JobRea	JobPro	JobHea	JobEnt	JobFod
midnight	.78	.77	11.68	−339.20	—	—	230.85	—	−431.29	15.88	—	—	−1.01	−13.25	—	—	—	20.26
1 a.m.	.69	.68	11.71	−255.41	—	—	177.18	—	−285.36	13.91	—	—	−1.01	−14.24	—	—	—	16.47
2 a.m.	.61	.60	9.49	−202.36	—	—	97.03	—	—	13.48	—	—	−1.11	−14.34	—	—	−2.49	15.23
3 a.m.	.57	.56	7.09	−176.59	—	—	70.29	—	—	10.51	—	—	−0.88	−11.85	—	—	−2.42	11.85
4 a.m.	.66	.65	17.14	−231.99	—	—	48.85	—	—	7.25	—	—	−0.56	−7.30	—	—	−1.30	7.81
5 a.m.	.76	.75	25.55	−321.28	0.29	—	50.97	—	−229.78	—	—	3.41	—	—	—	0.31	—	2.26
6 a.m.	.65	.65	13.51	−231.69	0.82	−177.08	—	0.13	—	—	—	—	—	—	1.53	0.80	2.41	—
7 a.m.	.74	.73	15.21	−302.89	1.50	−348.46	—	0.25	—	—	—	—	—	—	2.76	1.27	3.87	—
8 a.m.	.81	.80	29.08	−378.73	1.52	−395.80	—	0.32	—	—	—	—	—	—	4.26	1.46	4.80	—
9 a.m.	.83	.82	43.97	−404.27	1.26	−360.67	—	0.29	—	—	—	—	—	—	3.07	1.66	—	8.30
10 a.m.	.86	.86	57.19	−454.41	1.01	−308.34	—	0.24	—	—	—	—	—	15.05	—	1.63	—	8.87
11 a.m.	.89	.88	61.13	−508.46	—	—	—	0.17	—	—	—	—	—	11.95	2.04	1.16	3.95	6.55
noon	.90	.90	67.32	−538.85	—	—	—	0.17	—	—	7.23	—	—	—	3.46	1.02	3.54	7.39
1 p.m.	.91	.91	66.95	−564.52	—	−141.02	—	0.23	—	—	9.17	—	—	—	2.61	1.26	—	9.15
2 p.m.	.91	.91	63.60	−580.34	—	—	—	0.16	—	—	8.98	—	—	—	3.00	0.96	3.68	8.35
3 p.m.	.91	.91	64.45	−567.34	—	−154.67	—	0.22	—	—	8.03	—	—	10.51	—	1.37	—	9.42
4 p.m.	.91	.90	64.95	−554.60	—	−93.66	—	0.20	—	—	6.89	—	—	—	1.87	—	3.21	6.96
5 p.m.	.91	.90	63.58	−555.30	0.98	−288.01	—	0.24	—	—	8.46	—	—	—	2.31	—	—	10.73
6 p.m.	.92	.91	58.17	−583.47	—	—	—	0.22	—	—	6.59	—	−1.11	—	4.10	—	4.94	13.84
7 p.m.	.92	.91	55.86	−584.17	—	—	—	0.22	—	—	5.94	—	−1.24	—	4.04	—	5.30	16.35
8 p.m.	.91	.90	54.69	−557.68	—	−250.61	335.81	—	—	—	7.85	—	−1.19	—	3.32	—	—	18.86
9 p.m.	.90	.89	35.05	−528.50	0.93	−382.70	310.79	—	—	—	7.11	—	—	—	1.72	—	—	19.76
10 p.m.	.88	.88	21.45	−495.67	0.84	−363.06	289.95	—	—	—	5.46	—	—	—	—	0.84	—	22.69
11 p.m.	.84	.84	19.30	−428.32	0.77	−332.23	285.40	—	—	—	4.16	—	−0.93	—	—	—	—	21.11

NOTE: — = factor omitted from the model.

establishments are influential almost all day from pickup and drop-off trip generation coefficients, but the influence is relatively higher at breakfast time (7–9 a.m.), lunchtime (1 p.m.), and dinnertime (5–11:00 p.m.). These results provide 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 uses a large database of taxi trips with origins and destinations in New York City 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 with transit LOS and the *k* nearest neighbor algorithm, and a procedure was implemented to select important explanatory variables by using multiple linear regressions.

The TAT is mapped throughout 2,167 census tracts in New York City to compare with taxi demand, which clearly indicates the relationships between subway accessibility and taxi use. Taxi trips are more numerous in places where transit is more accessible. That 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—population, education, age, income, TAT, and total jobs—are shown to be influential in taxi trip generation modeling in New York City. 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 pickups or drop-offs is helpful in understanding how the relationships between taxi demand and those influential factors vary temporally and spatially.

ACKNOWLEDGMENT

This material is based on work supported by the U.S. Department of Transportation's University Transportation Centers Program.

REFERENCES

1. Schaller Consulting. *The New York City Taxicab Fact Book, 2006.* http://www.schallerconsult.com/taxi/taxifb.pdf. Accessed Nov. 1, 2013.
2. O'Neill, W.A., and E. Brown. *Transportation Research Circular E-C026: Long-Distance Trip Generation Modeling Using ATS.* TRB, National Research Council, Washington, D.C., 2001.
3. Ben-Edigbe, J., and R. Rahman. Multivariate School Travel Demand Regression Based on Trip Attraction. *International Journal of Social, Management, Economics and Business Engineering,* Vol. 42, No. 6, 2010, pp. 1169–1173.
4. Racca, D., and E.C. Ratledge. *Project Report for Factors That Affect and/or Can Alter Mode Choice.* Delaware Center for Transportation, University of Delaware, Newark, Del., 2004. http://udspace.udel.edu/handle/19716/1101. Accessed June 12, 2013.
5. Kumar, A., and D. Levinson. Specifying, Estimating, and Validating a New Trip Generation Model: Case Study of Montgomery County, Maryland. In *Transportation Research Record 1413,* TRB, National Research Council, Washington, D.C., 1993, pp. 107–113.
6. Schaller, B. A Regression Model of the Number of Taxicabs in U.S. Cities. *Journal of Public Transportation,* Vol. 8, No. 5, 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. Presented at 25th ARRB Conference—Shaping the Future: Linking Policy Research and Outcomes, Perth, Western Australia, ARRB Group, Melbourne, Australia, 2012.
9. Corpuz, G. Public Transport or Private Vehicle: Factors that Impact on Mode Choice. Presented at 30th Australasian Transportation Research Forum, Sydney, New South Wales, Australia, 2007.
10. Chang, Y.C. Factors Affecting Airport Access Mode Choice for Elderly Air Passengers. *Transportation Research Part E,* Vol. 57, Oct. 2013, pp. 105–115.
11. Schwanen, T., and P.L. Mokhtarian. What Affects Commute Mode Choice: Neighborhood Physical Structure or Preferences toward Neighborhoods? *Journal of Transport Geography,* Vol. 13, No. 1, 2005, pp. 83–99.
12. Ewing, R., W. Schroeer, and W. Greene. School Location and Student Travel: Analysis of Factors Affecting Mode Choice. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1895,* Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 55–63.
13. Schmöcker, J.D., M.A. Quddus, R.B. Noland, and M.G. Bell. Estimating Trip Generation of Elderly and Disabled People: Analysis of London Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1924,* Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 9–18.
14. Yazici, M.A., C. Kamga, and K.C. 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 the Transportation Research Board, No. 2308,* Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 83–95.
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, No. 390–398, 2005, pp. 390–398.
16. Loos, N. *Value Creation in Leveraged Buyouts: Analysis of Factors Driving Private Equity Investment Performance.* Deutscher Universitats-Verlag, Wiesbaden, Germany, 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, Los Angeles, Calif. http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm. Accessed Nov. 1, 2013.
18. Yan, X., and X.G. Su. *Linear Regression Analysis: Theory and Computing,* 1st ed. World Scientific Publishing Company, Singapore, 2009.
19. Fox, J. *Applied Regression Analysis and Generalized Linear Models,* 2nd ed. SAGE Publications, Inc., Thousand Oaks, Calif., 2008.
20. Davies, A. *Exhaustive Regression: An Exploration of Regression-Based Data Mining Techniques Using Super Computation.* Research Program on Forecasting. RPF Working Paper No. 2008-008. George Washington University, Washington, D.C. http://www.gwu.edu/~forcpgm/2008-008.pdf. Accessed July 27, 2013.
21. McLeod, A., and C. Xu. *Help bestglm: Best Subset GLM.* http://cran.r-project.org/web/packages/bestglm/vignettes/bestglm.pdf. Accessed July 26, 2013.
22. Elliott, G., A. Gargano, and A. Timmermann. *Complete Subset Regressions, 2012.* http://rady.ucsd.edu/docs/faculty/timmerman/subset-regression-April-25-2012.pdf. Accessed July 29, 2013.
23. Mallows, C.L. Some Comments on C_P. *Technometrics,* Vol. 15, 1975, No. 4, 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, No. 2, 2010, pp. 76–98.
25. Chatman, D.G. Does TOD Need the T? *Journal of the American Planning Association,* Vol. 79, No. 1, 2013, pp. 17–31.

The Transportation Demand Forecasting Committee peer-reviewed this paper.