EXPLORING THE SPATIAL DEPENDENCE AND SELECTION BIAS OF DOUBLE PARKING CITATIONS DATA

Jingqin Gao, M.Sc. (Corresponding author)
Graduate Research Assistant, C2SMART Center,
Department of Civil and Urban Engineering,
Tandon School of Engineering, New York University
Six MetroTech Center, 4th Floor, Brooklyn, NY 11201, USA
Tel: (646) 717-3652
E-mail: jingqin.gao@nyu.edu

Kun Xie, Ph.D.
Lecturer,
Department of Civil and Natural Resources Engineering
University of Canterbury
20 Kirkwood Ave, Christchurch 8041, New Zealand
E-mail: kun.xie@canterbury.ac.nz
Phone: +64-3-369-2707

Kaan Ozbay, Ph.D.
Professor & Director
C2SMART Center (A Tier 1 USDOT UTC)
Department of Civil and Urban Engineering & Center for Urban Science & Progress (CUSP)
Tandon School of Engineering
New York University
Six MetroTech Center, Room 404, Brooklyn, NY, 11201
http://c2smart.engineering.nyu.edu/
Tel (NYU CUE): 646.997.3691
Email: kaan.ozbay@nyu.edu

Word count: 5245 Texts + 5 Table + 4 Figures = 7495
Submission Date: November 15th, 2017

Abstract

Parking violation citations, often used to identify contributing factors to parking violation behaviors, is one of the most valuable datasets for traffic operation research. However, little has been done to examine its spatial dependence caused by location-specific differences in features such as traffic, land use, etc., and potential selection biases resulting from different levels and coverage of traffic enforcement. This study leveraged extensive data on double parking citations in Manhattan, New York in 2015, along with other datasets including land use, transportation and socio-demographic features. Moran's I statistics confirmed that double parking tickets were spatially correlated so that spatial lag and spatial error models were proposed to account for the spatial dependence of parking tickets to avoid biased estimates. To investigate whether selection bias exists in issuing tickets, we estimated the effects of parking ticket density and police precinct distance, when controlling for variables such as commercial area, truck activity, taxi demand, population, hotel and restaurant. Parking ticket density and police precinct distance were used as indicators of the enforcement levels and coverage and were found to be statistically significant. This indicated the existence of selection bias due to the heterogeneity in enforcement levels or coverage across different regions. Moreover, traffic enforcement units patrolling patterns revealed that the majority of the units have less than three daily patterns. These findings can assist proper usage of the citation data by suggesting researchers and agencies to consider spatial dependence as well as selection bias, and provide insights for parking violation management strategies.

Keywords: Parking tickets, Double parking, Spatial dependence, Selection bias, Traffic enforcement, Patrol patterns

INTRODUCTION AND MOTIVATION

For urban cities, parking can be extremely costly and imposes a substantial burden on drivers. In 2016, Americans are found to spend an average of 17 hours searching for parking at an estimated cost of 72 billion dollars in wasted time, fuel and emissions *(1)*. Other indirect parking pains include parking fines with an annual 2.6 billion dollars *(1)*. Parking violation management is nowadays an increasing concern of many mega cities such as New York City (NYC) due to the conflict between limited on-street parking supply and rigid daily demand of on-street parking. As a result, it often leads to illegal parking such as double parking that contributes significantly to traffic congestion, accidents, and transportation cost.

Every year, over 10 million parking violation tickets are issued in NYC and these citations become one of the valuable datasets for researchers to develop parking models that can help explore hotspots and to investigate the impacts of contributing factors. Conventional statistics relies on the assumption that each parking citation is independent. However, this assumption could be violated if spatial autocorrelation exists in the data, and thus spatial models are desired. On the other hand, with an increase in volume and variety of emerging data sources such as GPS-equipped vehicles, it enables more precise estimation of the effects of potential contributing factors by providing richer data for modeling *(2)*.

This also provides great opportunities for a deeper investigation of parking citation data, particularly on the potential underlying bias. Many studies in other disciplines have shown record-based data, such as crime data, is not a complete population of all occurrence, nor does it yield a representative random sample *(3)*. Although this point has not been fully proved in parking citations, intuitively, parking citations may underestimate the true occurrences of parking violations as most of such activities may not be ticketed. Correlation between total numbers of tickets may be proportional to different vehicle types (i.e. commercial/passenger vehicle) or the number of police that enforce a certain area. In many cases, the incidence of parking violations may be skewed to areas with high intensity enforcement, when in fact there may be a higher incidence in other areas. The fact that parking ticket data might not represent the actual population of parking violations accurately may due to the selection bias resulting from heterogeneous enforcement intensity. Selection bias in this case means certain areas are more likely to be ticketed by the enforcement officers than others. As such selection bias can be a problem when using ticket data, its existence and impact needs to be explored.

As an extension of two previous studies *(4, 5)*, this study leverages one year of geocoded parking tickets issued in 2015 in Manhattan, NYC with a focus on double parking, which is a unique phenomenon of urban cities. The objectives of the study are two-fold. The first goal is to question whether the spatial dependence exists in the double parking ticket data. If so, models without addressing the spatial dependence would lead to biased estimation. Second, this study aims to investigate whether selection bias exists in issuing parking tickets. We estimated the effects of two indictors of police enforcement intensity, parking ticket density and police precinct distance, when controlling for other variables such as land use, transportation, sociodemographic features, etc.

LITERATURE REVIEW

Various studies have been conducted for different cities using parking violation ticket data. Wang and Gogineni *(6)* conducted an empirical investigation of commercial vehicle parking violation behavior in NYC using one month of parking violation ticket data in May 2014. This study tried to identify the affecting factors that may influence the parking violation behavior of commercial

vehicles. Three models, Poisson regression, Negative binominal, and Zero-inflated negative binominal model (ZINB), were used to analyze parking violation frequency. Their results indicated land use and value, road type, on-street parking price, population and employment densities are related to the commercial parking tickets. However, although the study considered spatial distribution patterns of commercial vehicle parking violations, it did not account for potential spatial dependence structure and possible spatial bias while conducting regression analysis.

Wenneman *et. al (7)* established an ordinary least squares regression model to examine the relationship between the number of commercial vehicle parking tickets incurred in City of Toronto, the freight trip generation by establishments, and built environment factors represented by the parking supply. Although the final model claims to achieve an adjusted R-squared value of 0.68, this model did not consider the possible spatial dependence of the parking ticket data.

Kawamura *et. al (8)* conducted hot spots analysis using parking citations in Chicago. A regression model was developed to examine variables such as household income, sales from food services, etc. The results showed that two contrasting factors, concentrations of food businesses and stable neighborhoods can present problems for truck parking. One important contribution of this model is that it includes the log of the density of tickets issued to passenger vehicles and assumes that it reflects the number of parking enforcement and police officers in the area. The authors believed in the areas that are patrolled heavily, truck parking violations have a higher probability of being ticketed compared with areas with a low level of enforcement. A detailed summary of the above three studies can also be found in TABLE 1.

TABLE 1 Previous Studies on Parking Ticket Models

Study	Wang and Gogineni *(6)*	Kawamura *et. al (8)*	Wenneman *et. al (7)*
Dataset	One month (May 2014)	12-month (2011-2012)	12-month (2012)
Response Variable	Violation intensity (Violations/mile)	Log of truck ticket density	Parking citation density (citations/ zone)
Vehicle Type	Commercial Vehicles	Commercial Vehicles	Commercial Vehicles
Location	New York City, United States	Chicago, United States	Toronto, Canada
Methodology	Poisson regression model, Negative binominal model, and ZINB	Regression model with a combination of try-and-error and the backward elimination process	Ordinary least squares regression model
Key Explanatory Attributes	Land use types, Road types, population density, employment density, household, Parking prices	The number of establishments and employment, household income, population density, average rents and house values, retail / merchandize / food sale, vehicle availability, work trip mode shares, age, participation rate in online shopping, crime rate, language	The freight trip generation by establishments, number of loading zone spaces / loading bay doors, on-street parking spaces, on-street standing spaces, density of on-street standing spaces, number of surface lot spaces

Gao and Ozbay *(4)* studied the spatial distribution of NYC double parking violation records during July to October in 2014 for all vehicle types and commercial vehicles. The study showed

double parking had more violation records in commercial districts or mixed commercial/residential districts. Furthermore, field data collected by another study by Gao and Ozbay *(5)* confirmed that violation ticket data highly underestimated the real occurance of double parking violations. For example, a street block with a yearly record of 208 double parking tickets has more than 40 violations on a single day during 8AM to 9 AM.

Smith and Steif *(9)* used more than 1.4 million parking violation tickets in Philadelphia in 2013 to calculate the probability of at least one ticket being issued by hour and by street. They claimed that a more general probability cannot be estimated due to unobserved factors including the patrol patterns of traffic enforcement.

Although all of the above studies claimed that parking violation ticket data is very valuable, most studies have mainly focused on its relationship with land use, parking supply, and sociodemographic features *(6-8)*. There has been little discussion on its spatial dependence caused by location-specific differences (i.e. land use) and potential selection biases resulting from different traffic enforcement levels or coverage. Only a few studies *(5, 8, 9)* discussed the potential bias in ticket data, and only one study *(8)* attempted to analyze and quantify the bias due to parking enforcement and their model result found a strong association between "enforcement level" and parking violations tickets by trucks. Thus, there is a need to investigate spatial dependence and bias in the parking data, especially selection bias due to police precinct locations or patrolling patterns of traffic enforcement.

DATA PREPARATION

Parking Violation Tickets

Parking violation ticket data was released by New York City Department of Finance (NYCDOF) monthly and shared via NYC open data portal *(10)*. A batch geocoding program was developed based on U.S. Census Bureau Geocoding Services Web Application Programming Interface (API) *(11)* and Google Geocoding API *(12)* to convert the original address information from the parking ticket data into geographic coordinates. The geocoding rate can reach 95.3% after reformatting the address information and autocorrecting special cases with direction abbreviation or distance information (i.e. 200 feet W/O an intersection).

In total, 10,905,102 violation tickets were issued in the year 2015 for NYC. Jan. 27th (North American Blizzard), July 5th (the day after Independence Day) and December 27th (Weekend after Christmas) received the top lowest daily parking tickets in 2015. Among the total 10 million tickets, 633,050 out of them (6.3%) were committed by drivers who double parked. TABLE 2 lists descriptive statistics for all parking violations and double parking violations. The result lends support to the claim that unlike the general trend for all parking violations, commercial vehicles committed to a significant portion of the double parking violation tickets (45.2%). This may because of inadequate parking spaces for commercial vehicles in a highly dense urban network and the desire of the commercial vehicles to park as close as their delivery spot. The result also highlights that more than half of the double parking tickets were issued in Manhattan (58.8%), followed by Brooklyn (17.5%), Bronx (14.2%), Queens (7.8%) and Staten Island (0.2%). About 85%-90% of the tickets were issued during weekdays, and the hourly distribution pattern is consistent with work hours (FIGURE 1).

TABLE 2 Descriptive Statistics of All Parking Tickets/Double Parking Tickets

	All Parking Violations	Double Parking Violations
Total number of tickets in 2015	10,905,102	695,369
%Commercial vehicle tickets	19.4%	45.2%
%Passenger cars tickets	72.5%	47.3%
%Tickets issued in Manhattan	34.0%	58.8%
%Tickets issued in Brooklyn	20.7%	17.5%
%Tickets issued in Queens	18.3%	7.8%
%Tickets issued in Bronx	10.1%	14.2%
%Tickets issued in Staten Island	0.9%	0.2%
%Tickets issued during Weekday	85.7%	89.9%
%Tickets issued during Weekend	14.3%	11.1%

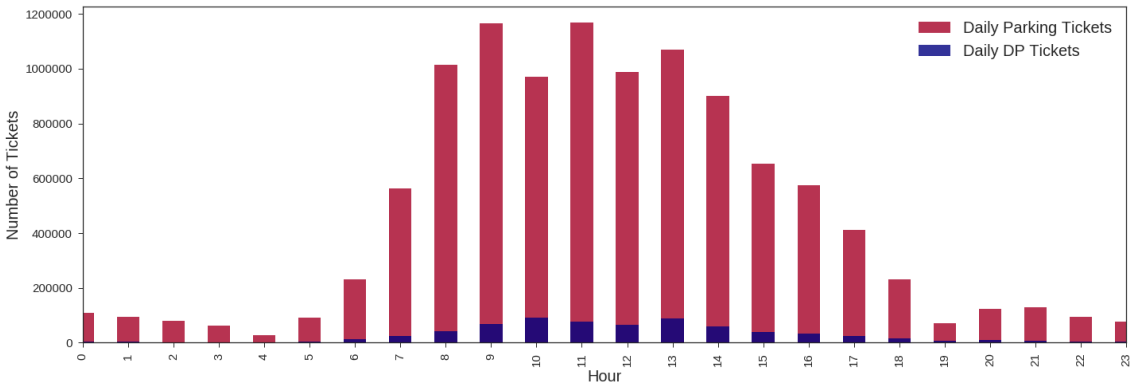


FIGURE 1 Hourly distribution of all parking/double parking tickets.

Geographical analysis units

The map of Manhattan is uniformly split into 6,204 equally sized grid cells (300 feet × 300 feet). The approximate width of a standard street block in Manhattan is close to 300 feet and the length of it (800~900 feet) can be integer multiples of 300 feet. Advantages of using equally sized grid cells as the basic geographical units include: 1) providing street-to-street resolution, 2) easing the bias from different analysis unit sizes (i.e. census tract), and 3) incorporating easily with data such as land-use features or taxi trips. Recent studies *(2, 13)* showed the implementation of cell-structured modeling framework in transportation research. Parking tickets along with sociodemographic, land use, transportation, and enforcement were aggregated for each cell using spatial analysis tools in ArcGIS *(14)*. FIGURE 2 demonstrates double parking ticket density at the grid-cell level in Manhattan in 2015.

FIGURE 2 Double parking ticket density at grid-cell level in Manhattan in 2015.

Land use, sociodemographic, transportation, and enforcement data

Various contributing factors to parking tickets have already been identified by previous studies. The most commonly used factors are land use and sociodemographic features like land use type and population *(6-8)*. This study obtained detailed and categorized land use zoning information for commercial, residential, mixed and park usage from New York City Department of Planning (NYCDCP) *(15)*. A Visual Basic for Applications (VBA) program was developed to compute the ratio for each zoning category in every grid cell *(2)*. Sociodemographic information based on 2011 census survey was retrieved from U.S. Census Bureau *(16)* and disaggregated into each grid cell.

In addition, Gao and Ozbay *(5)* and Kawamura *et. al (8)* pointed out that places of interest such as hotels or food sale places could be potential contributing factors for double parking. Therefore, hotel information from NYSDOT open data portal *(17)* and restaurant information collected by New York University researchers *(18)* was also utilized.

Road network features and traffic information *(6, 8)* are commonly used as well. In this study, taxi pick-ups and drop-offs obtained from NYC Taxi and Limousine Commission (TLC) *(19)* were used as estimates of traffic demand. Road network information like sidewalks, bike path, vehicle miles traveled (VMT) are obtained from NYCDCP, New York City Department of Transportation (NYCDOT), and New York State Department of Transportation (NYSDOT), respectively. A MapReduce program for expressing distributed and parallel computations was

developed to process the massive taxi trip records in 2015 (>20 Gigabytes) *(2)*. It should also be noted that the VMT for each grid cell was estimated from the length of the road segments in the grid cell and the average daily traffic of each roadway segment fall into the cell *(2)*. Public transit information such as bus and subway station Geographic Information System (GIS) data and the ridership for each subway station was generated using Metropolitan Transportation Authority (MTA)'s open data *(20)*.

There has been an inconclusive debate about whether traffic enforcement and patrol patterns have an impact on parking tickets. The number of tickets issued to a street block or an area may be biased by the level of enforcement and coverage. In order to investigate these factors, this study also collected the following data for each grid cell: 1) the density of tickets for all parking violations, and 2) the distance to the nearest police precinct station. The former is assumed to reflect the level of enforcement in the area. This variable was computed by counting the total number of parking tickets (all violation types) in each police precinct and then dividing this number by the precinct area, and then these values were assigned to each grid cell in the study region. The distance to the nearest police precinct station was computed for each grid cell. Police precinct station addresses were retrieved from New York Police Department (NYPD) website *(21)* and geocoded onto the map of our study area. TABLE 3 summarized the data descriptions and descriptive statistics.

TABLE 3 Data Descriptions and Descriptive Statistics (6,204 grid cells)

Variable	Description	Mean	SD
Dependent variable			
Double parking tickets	Annual double parking tickets	47.20	104.08
Land Use			
Commercial ratio	The ratio of commercial zone area to the whole area	0.29	0.40
Residential ratio	The ratio of residential zone area to the whole area	0.50	0.44
Mixed ratio	The ratio of mixed zone area to the whole area	0.06	0.22
Park ratio	The ratio of park area to the whole area	0.14	0.31
Hotel density	Number of hotels after spatial processing	0.04	0.23
Hotel Distance	Distance to the nearest hotel (mile)	0.95	1.45
Restaurant density	Number of restaurants after spatial processing	1.25	2.53
Enforcement			
Parking Ticket density	Number of parking tickets for all violation types per 1000 square feet	3.22	2.22
Police precinct distance	Distance to the nearest police precinct station (mile)	0.41	0.27
Sociodemographic			
Population	Total population	241.83	151.22
Population under 14	The population under 14 years	30.13	24.44
Population over 65	The population 65 years and over	32.10	25.70
Male	The population of males	113.86	70.73
Female	The population of females	127.95	82.21
Median age	Median age of population	1.58	0.99
Median income	Median income per household (10^3 $)	68.64	48.01

Employed	Number of the employed	129.51	87.47
Unemployed	Number of the unemployed	11.77	10.37
Transportation			
VMT	Annual vehicle miles traveled (10^3 veh. mile)	309.66	501.49
Truck ratio	The average ratio of truck flow to total flow	0.04	0.05
Subway ridership	Annual subway ridership after spatial processing (10^3)	245.76	392.20
Bus stop density	Number of bus stops after spatial processing	0.36	0.23
Sidewalk	Total length of sidewalks (mile)	0.07	0.07
Bike path	Total length of bike paths (mile)	0.02	0.03
Taxi pick-ups	Annual taxi pick-ups (10^3)	21.25	38.57
Taxi drop-offs	Annual taxi drop-offs (10^3)	20.30	29.99

METHODOLOGY

SPATIAL DEPENDENCE TEST

Moran's Test *I*

Spatial dependence test, such as Global Moran's I and local Moran's I *(22-24)* statistic tests, are widely used for measuring spatial autocorrelation *(25)*. If a spatial dependence exists in the data, commonly used assumptions for statistical models such as independent observation of the data will be violated, and the estimates from these models will be biased and inefficient. This type of bias may be corrected by using spatial models such as spatial lag or spatial error model.

Global Moran's I statistic proposed by Moran *(23)* illustrates an indication of "the degree of linear association between a vector of observed values and a weighted average of the neighboring values that underlies the specification of spatial autoregressive processes *(26)*". Local Moran's I is developed based on the assumption that global Moran's I is a summation of individual cross-products and computes a measure of spatial association for every observation *(24)*. The purpose of Local Moran's I statistics is to capture local patterns of spatial association that the global Moran's I may not observe.

Global Moran's I was used first to measure the spatial dependence of double parking tickets. A Moran's I value near 1.0 indicates clustering while a value near –1.0 indicates dispersion *(27)*. The z-score of Moran's and pseudo p-value *(28)* obtained from permutation test is used to assess the significance of Moran's I. A positive z-score suggests the distribution of the observations is spatially clustered *(29)* and a pseudo p-value less than 0.05 confirms that I is statistically significant at the confidence level of 95% *(30)*. More details about Global Moran's I practice can be found in a previous study by Xie *et. al (29)*.

GeoDa *(28)*, a spatial analysis software, was used to test whether double parking tickets were spatially correlated. The global Moran's I test was conducted using several weigh matrices (threshold distances, k-nearest neighbor) and 9,999 permutations were performed to compute the pseudo p-value. The results of global Moran's I test are presented in TABLE 4. All the pseudo p-values are found to be less than 0.05 using different weigh matrix. Since the result indicates a strong spatial autocorrelation on double parking ticket, it will lead to biased estimations and unreliable statistical inferences if such spatial dependence is ignored.

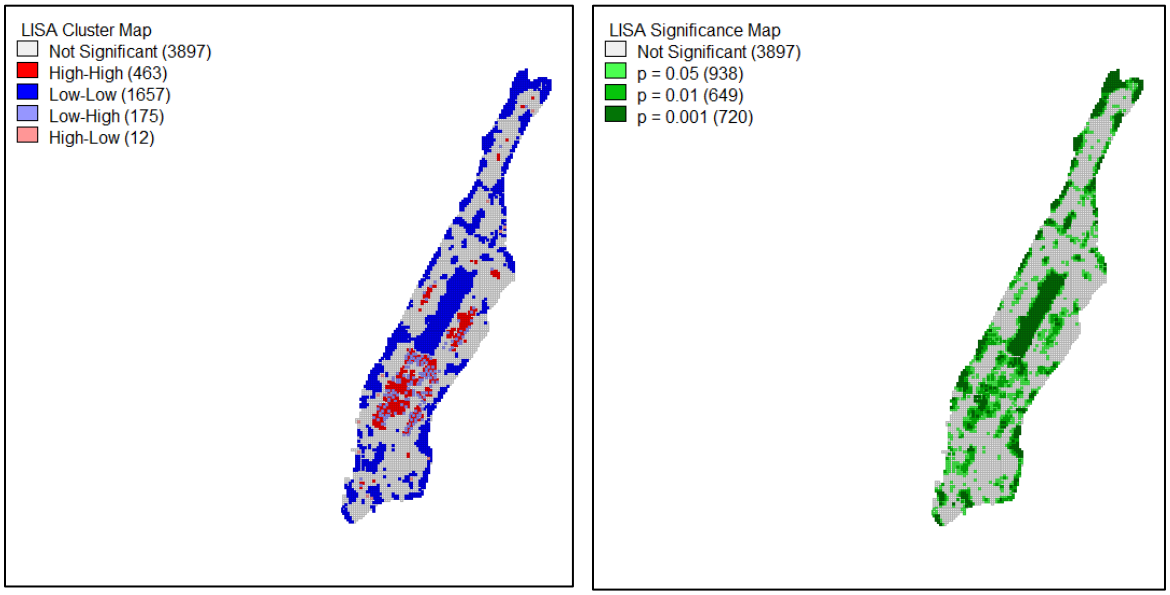
TABLE 4 Global Moran's I Test Results

Weight Matrix	I	E[I]	SD[I]	Z_i	Pseudo p-value
Threshold Distance-300 feet	0.3321	-0.0002	0.0090	36.9521	0.0001
Threshold Distance-800 feet	0.2581	-0.0002	0.0041	62.7201	0.0001
4-nearest neighbor	0.3318	-0.0002	0.0089	37.2234	0.0001
8-nearest neighbor	0.3262	-0.0002	0.0063	51.8185	0.0001

This study also conducted a local spatial autocorrelation analysis based on the Local Moran Local Indicators of Spatial Association (LISA) *(24)* statistics. While the global Moran's I measures spatial autocorrelation globally, LISA investigates individual locations and identifies hot spots and cold spots *(31)*. Local Moran's I for the observation z_i, z_j in cell i, j with weight matrix w_{ij} and N observations *(24)* can be computed as:

$$I_i = \frac{z_i}{(\frac{\sum_i z_i^2}{N})} \sum_j w_{ij} z_j \quad (1)$$

LISA cluster map (FIGURE 3(a)) is a choropleth map that illustrates high–high clusters, low–low clusters, high-low clusters, and low-high clusters. The high-high and low-low locations suggest clustering of similar high or low values, whereas the high-low and low-high clusters indicate spatial outliers *(32)*. In this study, low–low spatial clusters such as Central Park area is a cold spot, while high–high clusters such as Upper East Side can be regarded as a hotspot for double parking.



(a) LISA Cluster Map (b) LISA Significance Map

FIGURE 3 LISA cluster and significance map.

On the other hand, the LISA significance map is also a good indicator that shows the significance levels of the hots pots and cold spots identified by the cluster map *(32)*. FIGURE 3(b) shows the significance map using K-nearest neighbor (8 neighbors). Global and local Moran's I statistics confirmed that double parking tickets were spatially correlated so that spatial lag and

spatial error models in the next section were proposed to account for the spatial dependence of parking tickets to avoid biased estimates.

MODEL SPECIFICATION

We assume the spatial dependence identified from the previous section results from two aspects: 1) location-specific differences in features such as traffic, land use type, or population density, and 2) selection bias due to enforcement activities. The second aspect - selection bias due to enforcement activities is tested by introducing two new variables that were described in the data preparation section namely, parking ticket density and police precinct distance. Parking ticket density is assumed to be an indicator of level of enforcement and police precinct distance is assumed to be an indicator of enforcement coverage. To observe the effect of the selection bias - we keep the variables in the first aspect as control variables, and variables in the second aspect as experimental variables. The number of double parking ticket in each grid cell is used as dependence variable. Besides a standard linear model, a spatial error and a spatial lag model are examined since they are able to account for spatial dependence and unobserved spatial factors. The idea is that if traffic, land use, sociodemographic features and unobserved spatial factors are controlled, but the two experimental variables still show a strong association with the number of tickets issued, then a selection bias exists in the ticket data.

After diagnosing multicollinearity for the variables using variance inflation factors (VIF) *(33)*, the same set of control variables were selected for all three models so that valid model comparison can be compared. The selected control variables include commercial ratio, hotel distance, restaurant density, total populations, truck ratio, and taxi drop-offs.

A standard linear regression model assumes the error term to be independent and identically distributed with mean zero and constant variance. The method of ordinary least squares (OLS) estimation by minimizing the sum of squared prediction errors is widely used as a Best Linear Unbiased Estimator (BLUE) *(32)*. However, when we're predicting variables in space, the above assumption often does not hold as the errors can be spatially autocorrelated. In spatial analysis, there are two primary types of spatial dependence: spatial error and spatial lag (36). The former states that the spatial error terms across different spatial units are correlated. In the other words, omitted variables at one location can affect the dependent variable of itself and its neighboring locations (2, 37). The latter, spatial lag specification, allows spatial dependence through both spatial error correlation effects and spatial spillover effects (2, 37). The spatial error model can be expressed as follow:

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \mathbf{u} \tag{2}$$

$$\mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \tag{3}$$

where $\boldsymbol{\varepsilon}$ is a vector of independent normally distributed errors, while $\mathbf{u}$ is a vector of spatially autocorrelated errors, controlled by the term $\lambda\mathbf{W}\mathbf{u}$, where $\mathbf{W}$ is the spatial weight matrix, and the constant λ is the spatial autoregressive parameter that represents strength of the spatial autocorrelation.

The spatial lag model accounts for the spatial autocorrelation first, so that classic assumptions in equation (5) can still be kept for each error in error matrix ε considering them independent:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4}$$

where $\rho\mathbf{W}\mathbf{y}$ is a spatially lagged dependent variable, ρ is a spatial autoregressive parameter. The maximum likelihood estimation *(34)* was used to calibrate the spatial error model and the spatial lag model.

It should be noted that R^2 is no longer adequate to be used to measure the goodness-of-fit of spatial models because the residuals of spatial models are not independent of one another *(29)*. Therefore, alternative performance measurement based on likelihood estimation, such as AIC *(35)* or BIC *(36)* were used. The former introduces parameter number as a penalty term while the latter combines parameter number and sample size into the penalty term.

MODELING RESULTS, DISCUSSIONS AND LIMITATIONS

Coefficient estimates and statistic indicators, as well as model assessment measures are reported in TABLE 5. Considering street resolution and convergence time, the weight matrix with 300 feet distances was used in spatial lag and spatial error model.

Both models have lower AIC and BIC values, and that means they have substantial model improvement by considering spatial dependence. The autoregressive parameters ρ in the spatial lag model and λ in the spatial error model are highly significant, which provide confirmatory proofs that the double parking citations are spatially correlated. This finding suggested that spatial dependence should always be considered when conducting traffic operation research using parking ticket data. Ignorance of spatial dependence may lead to biased estimates.

Statistic indicator *p*-value was used to test the significance of experimental variables. Both experimental variables – parking ticket density and police precinct distance were found to be statistically significant at 95% level (p-values<0.05) in all three models. Parking ticket density, an indicator to level of enforcement, was found to have a positive impact on the number of double parking citations. The distance to the nearest police precinct was confirmed to be negatively associated with number of double parking tickets. In other words, the further away from the police station, the less likely a parking ticket would be issued. It is worth to point out that the spatial heterogeneity resulted from unobserved factors can be handled by the error term in our spatial models. As land use, transportation and sociodemographic features are controlled, and unobserved spatial factors are also taken into account in spatial models, both the experimental variables still show a strong correlation with the dependent variable. This indicated that parking tickets are closely related to enforcement intensity and further confirmed the existence of selection bias in parking data.

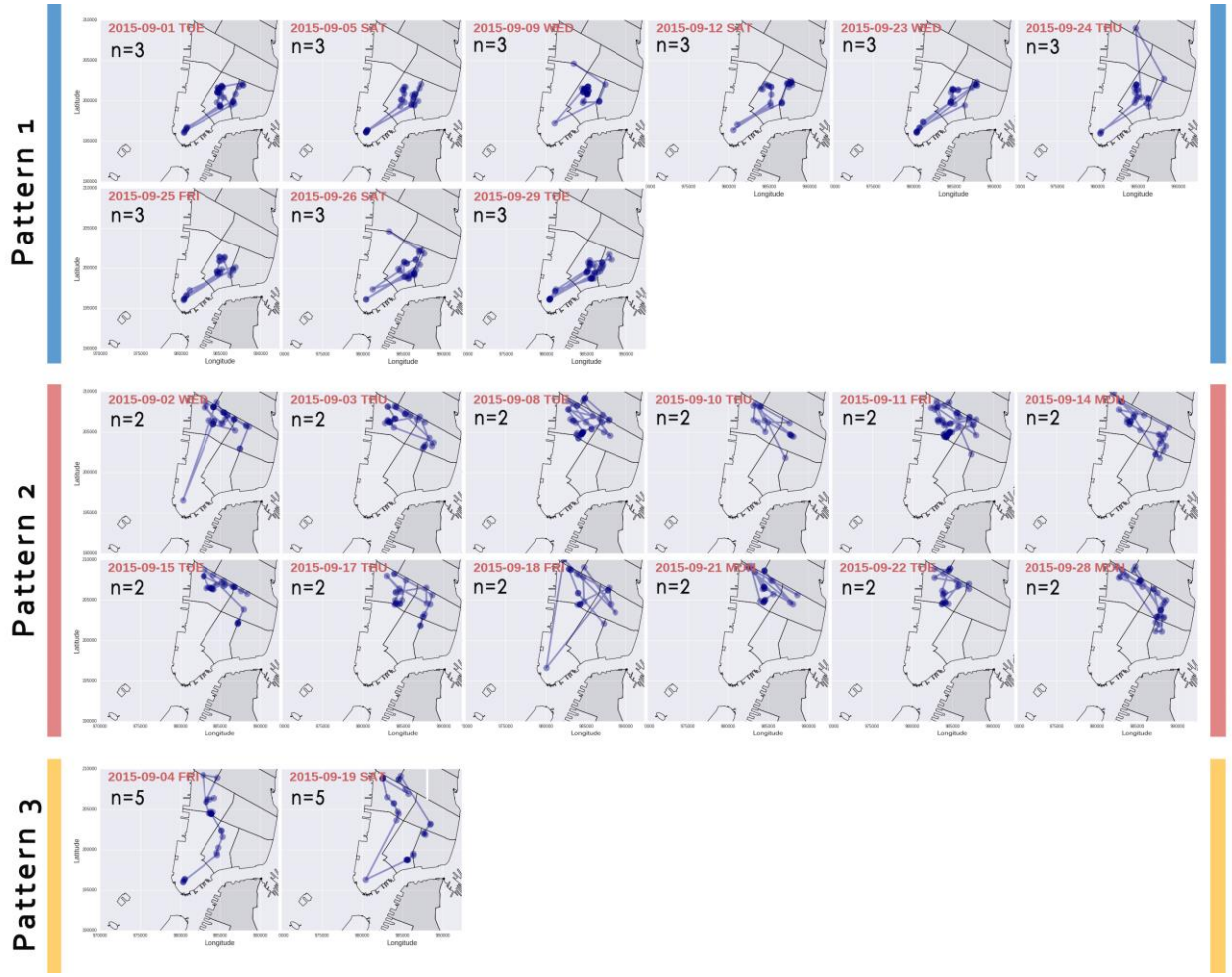
TABLE 5 Model Results and Assessment

Variables	Standard Model		Spatial Lag Model		Spatial Error Model	
	Coefficient	**p-Value**	**Coefficient**	**p-Value**	**Coefficient**	**p-Value**
Intercept	-12.712	<0.001	-14.472	<0.001	-11.501	0.006
Experimental variables						
Parking ticket density	5.313	<0.001	3.044	<0.001	5.199	<0.001
Police precinct distance	-18.433	<0.001	-14.665	0.002	-18.112	0.001
Control variables						
Commercial ratio	8.829	0.014	6.630	0.060	10.927	0.004
Hotel distance	6.606	<0.001	6.053	<0.001	6.592	<0.001
Restaurant density	7.316	<0.001	7.204	<0.001	7.311	<0.001
Total population	0.040	<0.001	0.037	<0.001	0.034	<0.001
Truck ratio	1.210	<0.001	1.225	<0.001	1.283	<0.001
Taxi drop-offs	0.886	<0.001	0.825	<0.001	0.862	<0.001
Autoregressive Parameter						
ρ			0.224	<0.001	-	-

λ		-	0.245	<0.001
Model Assessment				
R^2	0.225	0.259	0.263	
AIC	73683.300	73486.300	73469.900	
BIC	73743.800	73553.600	73530.500	

To further investigate potential characteristics of traffic enforcement, one month of geocoded parking ticket data for all types of parking violations (not only limited to double parking tickets) in September 2015 was utilized in this paper. "Issuer code" that represents unique code identifying issuing officer was used. For the same issuer code, the time sequence and location of issued parking tickets are assumed as a proxy for individual officer patrol route. In September 2015, 4,589 issuers issued parking tickets, among them, 3,108 issuers (68%) issued more than one ticket in the same time period. 43%, 24%, 18%, 10% of their daily patrol routes covers one, two, three and four police precincts, respectively. 5% of the daily patrol routes covers more than 5 precincts. Moreover, 50 issuers were randomly selected and their daily patrol routes were analyzed using GIS tools. The result shows that 80% of the issuers have less than three patrol patterns during the whole month, and 25% of the issuers prefer patrolling on avenues (major streets) to minor streets. FIGURE 4 demonstrates patrol pattern by the same issuer by day in the study month. Three different patterns were found according to the number of police precincts covered. In this study, we did not attempt to correct such bias by quantifying individual behaviors—a quite difficult endeavor at all levels. However, on the basis of the currently available data, considerable care must be taken when utilizing citation data due to selection bias identified in this study.

While the aim of the study is achieved, the current approach has limitations. Firstly, the variable "total ticket density" may not be completely true as a proxy of level of enforcement, especially if the study area is extended to other boroughs that have low parking citations in general. Secondly, the spatial models are not capable of accounting for unobserved, non-spatial factors. Unfortunately, unlike the long-studied crime or medical data model, parking models are limited to the available quantitative data. Future research efforts are needed in both data collection and methodology. Once more "unreported" parking data (i.e. identified from traffic cameras) becomes available, method such as synthetic population based approach *(3)* can be used to acquire a "ground truth" that containing a representative sample of parking violations.



*n= Number of police precincts covered by an issuer. A covered precinct is a precinct has more than two parking tickets issued per day by the same issuer.

FIGURE 4 Demonstration of patrol patterns of traffic enforcement.

CONCLUSIONS AND FUTURE WORK

This paper leveraged extensive data to investigate the spatial dependence of double parking citations via grid-cell-structured geographic framework and analyzed the existence of selection bias resulting from different levels of traffic enforcement or coverage of enforcement.

Massive empirical data from multiple sources, including parking tickets, land use types, place of interest, sociodemographic, enforcement, and transportation were collected, geo-processed, and cleaned. The descriptive analyses of the parking data show that although passenger vehicles are issued majority of parking tickets, double parking tickets does not follow this general trend. In fact, half of the double parking citations were given to commercial vehicles. The possible reasons of this finding can be 1) the lack of parking spaces for these commercial vehicles in a highly congested urban network 2) the desire of these commercial vehicles to park to the closest delivery spot even if parking space is not available to satisfy their delivery schedule 3) a combination of these two and other operational factors.

The global and local Moran's I statistics were in complete agreement that double parking citations were spatially dependent and such spatial dependence should not be neglected when using citation data. As a result, spatial lag and spatial error models were proposed to account for the spatial dependence of parking tickets to avoid biased estimates.

To investigate whether selection bias exists in issuing parking tickets, the effects of parking ticket density and police precinct distance were estimated while controlling for variables such as commercial area, truck activity, taxi demand, population, hotel and restaurant. Parking ticket density that is used as an indicator of the level of enforcement was found to have positive impact on the number of double parking tickets. When police precinct distance is used as an indicator of the enforcement coverage, it was found to be negatively correlated with the number of double parking tickets. Both of the experimental variables were found to be statistically significant, confirming the assumption that certain selection bias caused by enforcement intensity exists in the parking ticket data. Thus, it is recommended to be aware of this bias when using double parking ticket data to develop operational and tactical strategies to address the problem of double parking.

In addition, this study further contributed to the literature by investigating spatial bias potentially caused by patrol patterns of traffic enforcement. The result highlighted that majority of the issuers have less than three daily patrol patterns in the studied month and some of them have personal preference such as patrolling more heavily on major streets than minor streets. This further underlined the fact that considerable care must be taken when utilizing the citation data. Unfortunately, such bias is challenging to be quantified and corrected. Machine learning techniques such as unsupervised path clustering and more accurate data collection (i.e. installing GPS loggers on officer vehicles) may be applied as part of future research efforts to quantify the effectiveness of selection bias due to patrolling activities. Investigating passenger vehicle and commercial vehicle citation separately and examining all types of parking violations can also be part of the future work.

ACKNOWLEDGMENTS

The work in this paper is funded by C^{2}SMART, a Tier 1 University Transportation Center at New York University. The contents of this paper only reflect views of the authors who are responsible for the facts. The presented findings in the paper do not represent any official views or policies of any sponsoring agencies.

REFERENCES

1. Cookson, G. and B. Pishue. *The Impact of Parking Pain in the US, UK and Germany.* INRIX Research, 2017.
2. Xie, K., K. Ozbay, A. Kurkcu, and H. Yang, Analysis of traffic crashes involving pedestrians using big data: investigation of contributing factors and identification of hotspots. *Risk analysis*, 2017.
3. Lum, K. and W. Isaac, To predict and serve? *Significance 13(5)*, 2016, pp. 14-19.
4. Gao, J. and K. Ozbay, Modeling Double Parking Impacts on Urban Street. In: Proceedings of the Transportation Research Board 95th Annual Meeting, Washington, D.C., 2016.
5. Gao, J. and K. Ozbay, A Data-Driven Approach to Estimate Double Parking Events Using Machine Learning Techniques, 2017.
6. Wang, Q. and S. Gogineni, An Empirical Investigation of Commercial Vehicle Parking Violations in New York City. In: Proceedings of the Transportation Research Board 94th Annual Meeting, 2015.

7. Wenneman, A., M. Roorda, and K. Habib, Illegal Commercial Vehicle Parking, Parking Demand, and the Built Environment. In: Proceedings of the Proc., Canadian Transportation Research Forum, 2014.
8. Kawamura, K., P. Sriraj, H.R. Surat, and M. Menninger, Analysis of Built Environment Features and their Effects on Freight Activities. *Procedia-Social and Behavioral Sciences 125*, 2014, pp. 28-35.
9. Smith, T. and K. Steif, Street by hour probability of getting a parking violation in Philadelphia, Accessed, July, 2017.
10. New York City, NYC Open Data, https://opendata.cityofnewyork.us/, Accessed May, 2017.
11. U.S. Census Bureau. *Geocoding Services Web Application Programming Interface (API).* 2017.
12. Google, Geocoding API, developers.google.com/maps/documentation/geocoding, Accessed July, 2017.
13. Xie, K., K. Ozbay, Y. Zhu, and H. Yang, Evacuation Zone Modeling under Climate Change: A Data-Driven Method. *Journal of Infrastructure Systems 0(0)*.
14. ESRI. *ArcGIS Desktop Release 10.* 2011.
15. NYCDCP, www.nyc.gov/planning, Accessed June, 2017.
16. United States Census Bureau, https://www.census.gov/en.html, Accessed May, 2017.
17. New York State, I Love NY - Places to Stay Map, https://data.ny.gov/dataset/I-Love-NY-Places-to-Stay-Map/3ynz-ayvg/data, Accessed May, 2017.
18. Thompson, M., Grade "A" Restaurants in Manhattan. 2015.
19. NYC Taxi& Limousine Commision, Trip Records, http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml, Accessed Jan, 2017.
20. MTA, Turnstile Data, http://web.mta.info/developers/turnstile.html, Accessed June, 2017.
21. New York Police Department, Precinct, http://www1.nyc.gov/site/nypd/bureaus/patrol/precincts-landing.page, Accessed May, 2017.
22. Moran, P.A., Notes on continuous stochastic phenomena. *Biometrika 37(1/2)*, 1950, pp. 17-23.
23. Moran, P.A., The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological) 10(2)*, 1948, pp. 243-251.
24. Anselin, L., Local indicators of spatial association—LISA. *Geographical analysis 27(2)*, 1995, pp. 93-115.
25. Tiefelsdorf, M., The saddlepoint approximation of Moran's I's and local Moran's Ii's reference distributions and their numerical evaluation. *Geographical Analysis 34(3)*, 2002, pp. 187-206.
26. Anselin, L., *The Moran scatterplot as an ESDA tool to assess local instability in spatial association*, Regional Research Institute, West Virginia University Morgantown, WV, 1993.
27. ESRI. *Spatial Autocorrelation (Morans I) (Spatial Statistics).*
28. Anselin, L., I. Syabri, and Y. Kho, GeoDa: an introduction to spatial data analysis. *Geographical analysis 38(1)*, 2006, pp. 5-22.
29. Xie, K., K. Ozbay, and H. Yang, Spatial analysis of highway incident durations in the context of Hurricane Sandy. *Accident Analysis & Prevention 74*, 2015, pp. 77-86.
30. Goodchild, M.F., *Spatial autocorrelation*, Geo Books, 1986.
31. Zhang, C., L. Luo, W. Xu, and V. Ledwith, Use of local Moran's I and GIS to identify pollution hotspots of Pb in urban soils of Galway, Ireland. *Science of the total environment 398(1)*, 2008, pp. 212-221.
32. Anselin, L., GeoDa 0.9 user's guide. *Urbana 51*, 2003, pp. 61801.
33. O'brien, R.M., A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity 41(5)*, 2007, pp. 673-690.

34. Ord, K., Estimation methods for models of spatial interaction. *Journal of the American Statistical Association 70(349),* 1975, pp. 120-126.
35. Akaike, H., A new look at the statistical model identification. *IEEE transactions on automatic control 19(6),* 1974, pp. 716-723.
36. Schwarz, G., Estimating the dimension of a model. *The annals of statistics 6(2),* 1978, pp. 461-464.