Article

Spatiotemporal Varying Effects of Built Environment on Taxi and Ride-Hailing Ridership in New York City

Xinxin Zhang [1], Bo Huang [2],* and Shunzhi Zhu [1]

[1] College of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China; zhangxinxin@xmut.edu.cn (X.Z.); szzhu@xmut.edu.cn (S.Z.)

[2] Department of Geography and Resource Management and Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Hong Kong 999077, China

* Correspondence: bohuang@cuhk.edu.hk

check for updates

Abstract: The rapid growth of transportation network companies (TNCs) has reshaped the traditional taxi market in many modern cities around the world. This study aims to explore the spatiotemporal variations of built environment on traditional taxis (TTs) and TNC. Considering the heterogeneity of ridership distribution in spatial and temporal aspects, we implemented a geographically and temporally weighted regression (GTWR) model, which was improved by parallel computing technology, to efficiently evaluate the effects of local influencing factors on the monthly ridership distribution for both modes at each taxi zone. A case study was implemented in New York City (NYC) using 659 million pick-up points recorded by TT and TNC from 2015 to 2017. Fourteen influencing factors from four groups, including weather, land use, socioeconomic and transportation, are selected as independent variables. The modeling results show that the improved parallel-based GTWR model can achieve better fitting results than the ordinary least squares (OLS) model, and it is more efficient for big datasets. The coefficients of the influencing variables further indicate that TNC has become more convenient for passengers in snowy weather, while TT is more concentrated at the locations close to public transportation. Moreover, the socioeconomic properties are the most important factors that caused the difference of spatiotemporal patterns. For example, passengers with higher education/income are more inclined to select TT in the western of NYC, while vehicle ownership promotes the utility of TNC in the middle of NYC. These findings can provide scientific insights and a basis for transportation departments and companies to make rational and effective use of existing resources.

Keywords: geographically and temporally weighted regression; taxi; Uber; spatiotemporal analysis

1. Introduction

With the popularity of mobile phone usage, transportation network companies (TNCs) that offer app-based services, such as Uber, DiDi, and Lyft, claim to provide stability and convenience with peer-to-peer (p2p) processes that connect passengers and private drivers on-line and in real-time [1]. As an emerging form of transportation based on network and mobile technology, the analysis of TNC ridership has become a hot topic in urban transportation research. Much evidence has shown that the rapid development of TNC has had a huge impact on the traditional taxi (TT), leading the taxi industry to experience significant losses in terms of market share, revenue, labor power and facility [2]. This is particularly obvious in large modern cities such as New York City (NYC), where the annual taxi load decreased from 145 million in 2015 to 113 million in 2017, decreasing nearly 23% in three years. In contrast, the ridership by TNCs increased from 37 million to 110 million. The reduction in the market share of the taxi industry will inevitably cause a decline in the income of taxi drivers and the

compression of the taxi business scale, leading to economic difficulties and even the bankruptcy of taxi companies. In May 2013, although the price of a yellow car's license plate in NYC had been cut in half, the licenses of many taxi company vehicles were idle because of the lack of new drivers [3].

Nevertheless, many researchers insist that it is premature to announce the inevitable demise of the taxi industry based on the current success of TNC. For example, Wang et al. reported that the success of TNCs lies in an aggressive but unsustainable price subsidy strategy [4]. Cramer and Krueger's study [5] observed that most trips on TNC are concentrated in daily traffic peak periods. Regarding off-peak periods, traditional taxis still account for a large proportion of transportation and thus cannot be replaced. Furthermore, according to the statistical results from [6], the average number of working hours per week of Uber drivers was approximately half that of many taxi drivers in the U.S.

Regardless of these debates, it is indisputable that the taxi industry is currently facing a huge challenge and competition from the TNC in many aspects. Therefore, analysis of the differentiation of these two modes, such as the characteristics of the target passengers and travel pattern, is conducive to a better understanding of the competitive relationships between them. However, as all these differentiations are not uniform within a city and are driven by diverse factors, the widely used global statistical models are limited to incorporate the significance of spatiotemporal heterogeneity and autocorrelation. The spatiotemporal analysis between taxi/TNC ridership and the built environment is still an open issue.

This paper presents the results of our research utilizing an improved GTWR model based on parallel computation to efficiently explore the spatiotemporal relationships between TT/TNC and the built environment in NYC, where about 659 million trips occurred from 2015 to 2017. The rest of this paper is arranged as follows. Section 2 provides a brief review of the relevant research progress, and Section 3 presents the details of the parallel-based GTWR model adopted in this study. Section 4 introduces the related dataset and describes how the data were processed. Section 5 mainly analyzes the model accuracy and findings. Section 6 discusses the spatiotemporal patterns between taxi and TNC. The last section elaborates upon the conclusions of this paper, as well as future research directions.

2. Related Literature

Taxis have historically comprised a far lower share and geographical coverage of urban transportation than other transport modes, such as buses and subways; therefore, there are many lesser extensive studies on taxis than on other transport modes. In general, researchers have found taxis to be both complements and substitutes for public transit [2]. Despite their small share in urban transportation, taxis fill a critical gap by providing mobility service and all-day operation, which are not available in other transportation modes. More importantly, with the popularization of GPS auto-collection devices, the spatiotemporal characteristics of ridership and trajectory by taxis provide a valuable reference for mining the travel patterns of citizens and for traffic optimization [7]. Therefore, the spatiotemporal analysis of taxis has become a research hotspot in recent years.

Early research on taxis mainly focused on market demand components based on the inherent attributes of the taxi industry, such as price, tips, labor costs, and other factors [8]. Because the measurement of cost, waiting time, and convenience is usually derived from investigations or relevant departments, those data are biased and lack objectivity. With the GPS devices carried by taxis, the spatiotemporal data of taxis can be tracked and collected in real-time. These data have the advantage of spatial-temporal characteristics than previous data and can integrate with external geographic factors, such as land use [9] and weather [10,11]. For example, Liu et al. used GPS data of taxi and urban land use factors to identify 'source-sink areas' in Shanghai [12]. Nevertheless, previous studies mostly adopted the ordinary least squares (OLS) method [13,14]. In the OLS model, the aggregated pickup (PU) and drop-off (DO) locations of taxis are used as dependent variables, and the relevant influencing factors, such as weather and land use, are selected as independent variables. Given spatial autocorrelation and heterogeneity exists in the distribution of PU and DO locations for

TT and TNC, the precondition of the OLS model that the observations should be independent of each other is difficult to satisfy.

To address this issue, Fotheringham et al. proposed a local regression model called Geographically Weighted Regression (GWR) [15], which improves the accuracy of regression results by constructing a local spatial weight matrix for estimating variation in space. Furthermore, the GWR model extends the traditional regression framework by allowing parameter estimates to vary in space and is therefore capable to capture local effects. The GWR model has been widely applied in transit ridership analysis [16,17]. For example, based on NYC's taxi data, Qian et al. [18] used the GWR model to analyze the relationship between taxi locations and urban environmental factors. The results show that the GWR model can provide better model accuracy and interpretation than the OLS model. One of the remaining problems is that the GWR model only obtains related variable coefficients in the spatial dimension. While dealing with time series datasets, those data often need to be aggregated or separated based on their timestamps, thereby ignoring the fact that the distribution of taxis or TNCs varies with different scales of time. Recently, scholars have put forward many improved strategies to account for both temporal and spatial variability, such as the GWR-TS [19] and linear mixed effect (LME) + GWR models [20]; still, these models are generally based on the two-stage least squares regression [21], first fitting the temporal effect using the LME model and then evaluating the spatial heterogeneity effects with the GWR model. Those models cannot simultaneously consider temporal and spatial effects.

To simultaneously model temporal and spatial effects, Huang et al. proposed an improved GWR-based model, named Geographical and Temporal Weighted Regression (GTWR) [22], which is thought to design simultaneous spatial and temporal weighting. Thus, the GTWR model can reflect continuous variations for each location at each time. The initial implementation of the GTWR model was carried out for house-price estimation, and the results showed that the accuracy of the GTWR model was superior to that of the OLS and GWR models. Recently, the GTWR model has been extended in many fields, such as air quality [23] and environmental research [24]. Moreover, some scholars have put forward improved GTWR schemes successively. For example, Wu et al. proposed an improved model, known as the Geographically and Temporally Weighted Autoregressive (GTWAR) model, to estimate spatial autocorrelations [25], and Du et al. proposed a Geographically and circle-Temporally Weighted regression (GcTWR) model for enhancing the seasonal cycle of long-term observed data [26].

The above research fully shows that the GTWR model has great advantages in spatiotemporal modeling. Ma et al. applied the GTWR model to public transit and achieved good modeling results [27]. Zhang et al. also adopted the GTWR model to taxi ridership analysis and achieved a similar conclusion [28]. Nevertheless, due to the fact that the spatiotemporal nonstationarity of taxis is more complicated than other modes of transit such as buses that have preset routes, previous studies have generally been limited to taxis or TNC separately, and few studies take into consideration the difference between taxis and TNC. Research on TNC remains relatively scarce, although its data structure is similar to that of taxis. Thus, applying the GTWR model for simultaneous analysis of both taxis and TNC is still an unsolved issue.

3. Methodology

In this section, we briefly review the basic framework for the GTWR model and how to determine the parameters of the GTWR model. Then, we propose a parallel computing scheme to improve the efficiency of the GTWR model and apply the model to ridership modeling.

3.1. The Basic Framework of the GTWR Model

The GWR-based model is a local-based spatially varying coefficient regression algorithm that extends the OLS model by adopting local parameters to be estimated. It is capable of significantly improving the estimation accuracy of spatial data, especially for those areas with complex spatial nonstationarity. On this foundation, Huang et al. [22] proposed a GTWR model focusing on

spatiotemporal kernel function definition and spatiotemporal bandwidth optimization, which can simultaneously address spatial and temporal nonstationarity issues. Assuming that the observation of taxi ridership is denoted as Y_i, where i ($I = 1, 2, \ldots, n$) represents a spatial unit, such as traffic analysis zone (TAZ), thus the GTWR model can be mathematically expressed as follows:

$$Y_i = \beta_0(u_i, v_i, t_i) + \sum_k \beta_k(u_i, v_i, t_i)X_{ki} + \varepsilon_i, \tag{1}$$

where (u_i, v_i, t_i) represents the center coordinates of TAZ i in a spatial location (u_i, v_i) at time t_i; β_0 is the intercept value; $\beta_k(u_i, v_i, t_i)$ denotes the slope for each independent variable X_{ki}; and ε_i is the random error. The variables X_{ki} refer to the influencing factors that improve the associations between ridership and urban environmental factors, such as weather, land use, socioeconomic, and transport condition.

For a given dataset, a locally weighted least squares method is usually employed to estimate the intercept of β_0, as well as the slopes β_k for each variable. The GTWR models assume that the closer to point i in the space-time coordinate system, the greater the weight of the measurements in predicting β_k will be. Thus, the coefficients of $\hat{\beta} = (\beta_0, \beta_1, .., \beta_k)^T$ can be estimated by:

$$\hat{\beta}(u_i, v_i, t_i) = [\mathbf{X}^{\mathbf{T}}\mathbf{W}(u_i, v_i, t_i)\mathbf{X}]^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{W}(u_i, v_i, t_i)\mathbf{Y}, \tag{2}$$

where $\mathbf{X}$ is the $n{\times}(k{+}1)$ matrix of input variables. $\mathbf{Y}$ is the n-dimensional vector of the output variables. The space-time weight matrix $\mathbf{W}(u_i, v_i, t_i)$ is an $n \times n$ weighting matrix to measure the importance of point i to the estimated point j for both space and time. The Gaussian function is one of the most commonly used weight function:

$$W_{ij} = \exp(-\frac{d_{ij}{}^2}{h^2}), \tag{3}$$

where d_{ij} denotes a spatiotemporal distance between points i and j, and h is a nonnegative parameter that presents a decay of influence with distance. By combining the temporal distance d^T with the spatial distance d^S, the spatiotemporal distance can be expressed as:

$$d^{ST} = d^s \otimes d^T, \tag{4}$$

where '$\otimes$' can represent different types of operators. In this study, the '+' as the combination operator was selected to calculate the total spatiotemporal distance. With respect to the different scale effects of space and time, an ellipsoidal coordinate system was constructed to measure the spatiotemporal distance between each regressive point and the surrounding points [29]. The spatiotemporal distance between taxi ridership can thus be expressed as the linear weighting combination indicated below:

$$(d_{ij}^{ST})^2 = \lambda[(u_i - u_j)^2 + (v_i - v_j)^2] + \mu(t_i - t_j)^2, \tag{5}$$

where t_i and t_j denote the observed time of point i and j. λ and μ are the weights for balancing the influences of differing units between space and time variability. The weight matrix is constructed by using the Gaussian distance decay-based functions and Euclidean distance:

$$\begin{aligned} W_{ij} &= \exp[-\frac{(d_{ij}^{ST})^2}{h_{ST}{}^2}] \\ &= \exp\{-\frac{[(u_i-u_j)^2+(v_i-v_j)^2]+\tau(t_i-t_j)^2}{h_s{}^2}\} \end{aligned} \tag{6}$$

where the parameter τ stands for the non-negative parameter of scale factor calculated by μ/λ ($\lambda \neq 0$). h_{ST} is a positive parameter named the spatiotemporal bandwidth. Thus, if the spatiotemporal bandwidth and scale factor are determined, the weight matrix $W(u_i, v_i, t_i)$ and $\hat{\beta}(u_i, v_i, t_i)$ can be obtained.

The adjustment parameters of h_{ST} and τ can be acquired either utilizing a cross-validation (CV) process via minimization in terms of the corrected Akaike information criterion (AIC_c) [30] as follows:

$$\text{CVRSS}(h) = \sum_i [y_i - y_{\neq 1}(h)]^2,$$ (7)

$$AIC_c(h) = n \log\left(\frac{RSS(h)}{n}\right) + n\log(2\pi) + n\left(\frac{n + tr(H(h))}{n - 2 - tr(H(h))}\right),$$ (8)

where $y_{\neq 1}(h_s)$ indicates the predicted value y_i from the GTWR model with a bandwidth of h. Therefore, the selection of the optimum h can be acquired through plotting $CV(h)$ against the parameter h. In Equation (8), RSS is the residual sum of squares, and $tr(H(h))$ is the trace of the hat matrix $H(h)$.

3.2. Implementation of GTWR for Ridership Analysis

Figure 1 presents a flowchart of the implementation using GTWR for ridership analysis. Before constructing the GTWR model for taxi ridership analysis, the observed spatial unit and temporal resolution must be determined first. The spatial unit is generally related to the geographic extent of the study area, which can be divided by administrative regions or a regular cell. Due to the limitation of TNC data obtained from NYC, the spatial unit adopted in this study was based on TAZ rather than Zip Code Tabulation Areas (ZCTA). In terms of the temporal resolution, different resolutions, such as yearly, monthly, daily, and hourly scales, can be adopted. Since the dataset we adopted was from 2015 to 2017, the month was considered an appropriate minimum unit of time to reduce the cost of computation. Using the same dataset, the OLS and GTWR models were respectively applied to estimating the globe and local coefficients for both modes and their relationship with the urban architecture environment.
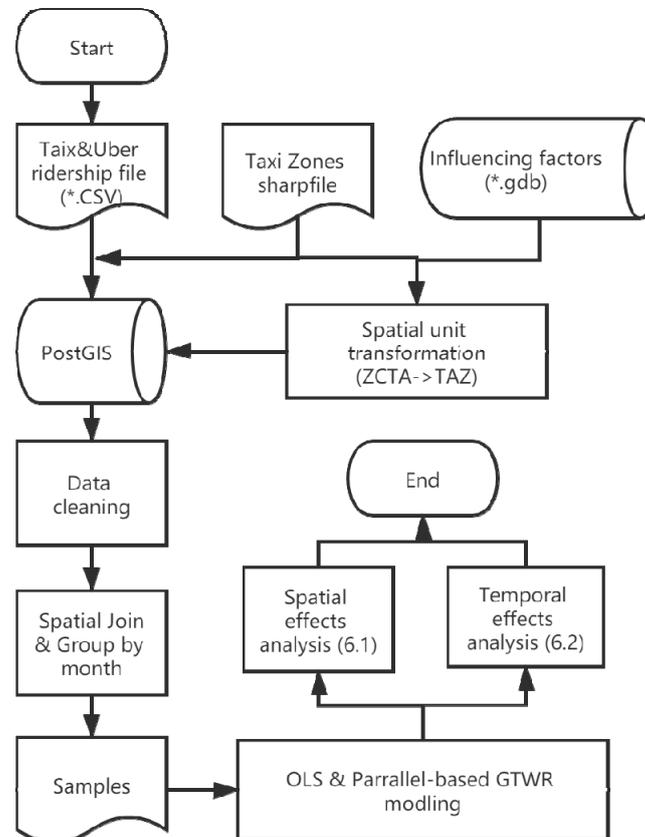

Figure 1. The flowchart of the implementation using GTWR for ridership analysis.

To quantitatively evaluate the spatiotemporal variation of ridership for taxis and TNC, three variables, including the ridership for two types of TT (yellow + green), the ridership of TNC and the proportion of TNC (PoT = TNC/(TT+TNC)), were selected as dependent variables. With respect to independent variables, we extracted four groups of explanatory variables from multiple open datasets. More details about raw data processing can be found in Section 4.

Several aspects need to be adjusted when applying the GTWR model to taxi ridership analysis. Firstly, compared with the fixed kernel function, the adaptive kernel function can adjust bandwidths according to the density of data points. Thus, it might be a more reasonable way to obtain the weights W_{ij} for the irregular sharp of TAZs. For simplicity, we use the q-nearest neighbors based on the following modified bi-square function:

$$W_{ij} = \begin{cases} \left[1 - (d_{ij}/h_i)^2\right]^2, & \text{if } d_{ij} < h_i \\ 0, & \text{otherwise} \end{cases}, \tag{9}$$

where h_i stands for the different bandwidths, which express the q nearest neighbors to consider in the estimation of regression at location i. Thus, the adjustment parameter of fixed bandwidth h_{ST} is replaced with the number of nearest neighborhood points q. Note that the q should be constrained to $q \geq 40$, otherwise the model will suffer over-fitting problem [25].

The computation of the GTWR model is intensive because each sample uses an adaptive type of bandwidth, which leads to $(t^*(n-1)^n)$ combinations of possible values that must be computed for the optimal bandwidth [15]. The computing time will exponentially increase as the number of samples and timestamps increases by, for example, using grid-based data as the spatial unit or constructing the daily GTWR model based on several years of data. An optimized modeling approach is needed to reduce computation consumption. In particular, we employed parallel computing to break down the computational loops of optimal parameter selection into independent parts with different values of q and τ. These parts can be executed simultaneously by multiple processors communicating via shared memory, the results of which are combined upon completion as part of the overall algorithm. Thus, the optimal values of q and τ could be efficient found. According to the principle of GTWR, the main computing power is consumed by iteration for obtaining both adjustment parameters. Since the process of finding the optimal adjustment parameters is independent for each iteration, it is suitable for parallel programming. The GTWR algorithm adopted in this study is programmed based on Matlab®, which provides a function called *fminbnd* for obtaining the optimal value q and τ. The process of parallel computing can be realized through a loop process, namely *parfor* [31]. The platform for efficiency comparison is based on an Inter® i7-4790 CPU, which has four cores for parallel computing. Figure 2 shows the flow chart of GTWR model using parallel computing. To verify the robustness of the results, different proportions of training samples were randomly selected for the GTWR model, while the remaining samples were used for verification.
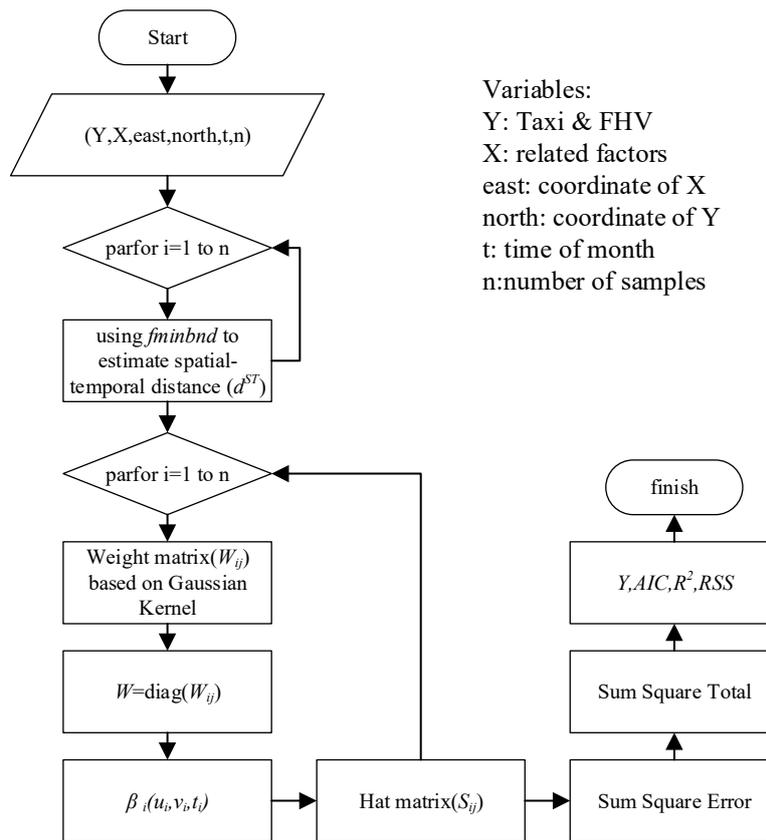
Figure 2. The flow chart of GTWR model using parallel computing.

4. Data Preparation

4.1. Study Area

The study area of NYC consists of five boroughs, including the Bronx, Brooklyn, Staten Island, Queens, and Manhattan. As shown in Figure 3, NYC is divided into 263 taxi zones, including three airport zones, 55 yellow zones (only yellow cabs are allowed to pick up passengers) in Manhattan, and 205 borough zones, in which both types of taxis are allowed to operate. TNCs are allowed in all areas. Previous scholars' research [32] reported that 95% of yellow passengers are concentrated in the Manhattan area, indicating that there were obvious imbalances in terms of spatial distribution and emphasizing the need to establish the GTWR model.
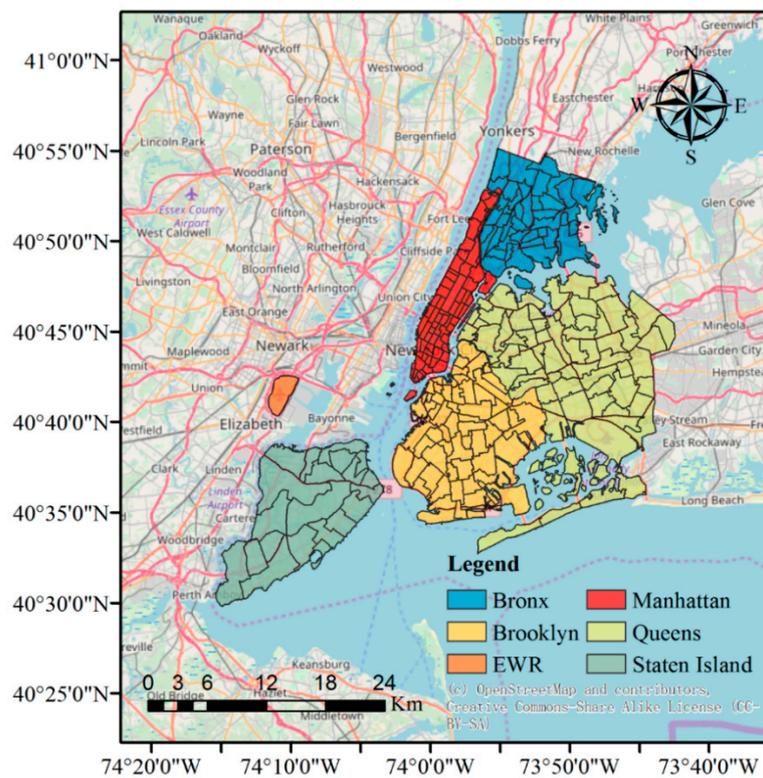
Figure 3. Taxi zones of NYC (263 in total, classified in five boroughs).

4.2. Taxis and TNC Data

The raw data from 2015 to 2017 were download from the NYC Taxi and Limousine Commission (TLC, available at http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml). The department provided three types of data from 2009 to 2018, including two types of traditional taxis (yellow and green) and TNC data, in CSV format. Each trip record on taxis included the PU and DO timestamps and locations, number of passengers, travel time, travel distance, and price attributes. However, instead of Pick-Up (PU) and Drop-Off (DO) points, the TNC trip data that was public since 2015 only provided the taxi zones due to privacy protection. Currently, NYC has three typical taxi modes, including: (1) yellow taxi serving anywhere within the city boundary; (2) green taxis (Boro taxi) serving only serving city remote areas except for two airports; (3) TNC serving the same extent as yellow taxis. Table 1 provides more details about the summary statistics of the three types of taxis, respectively. The total number of recorded PUs by yellow cars was 390 million, the total number of recorded PUs by TNCs was 212 million, and the total number of recorded PUs by green cars was only 57 million.

Table 1. Statistical description of two types of taxis and TNC data.

Type	2015	2016	2017	Total
Yellow	146,112,989	131,165,043	113,496,706	390,774,738
	72.28%	59.80%	47.46%	59.15%
TNC	36,910,806	69,131,726	106,676,500	212,719,032
	18.26%	31.52%	44.60%	32.20%
Green	19,116,598	19,054,688	18,990,815	57,162,101
	9.46%	8.69%	7.94%	8.65%
Total	202,140,393	219,351,457	239,164,021	660,655,871

Due to the limitation of data security, the downloaded TNC data only included the timestamp of PU and DO and the TAZ's ID where both coordinates were located. Therefore, to ensure the unity of

spatial reference, the taxi zone defined by TLC was adopted as the basic spatial unit, and number of months was selected as the temporal unit.

After determining the spatiotemporal unit, i.e., each observation represented the total number of ridership at one taxi zone in a certain month, the dependent variables of monthly ridership were derived based on the spatial and temporal aggregations of each trip. First, the raw data were imported into the PostGIS spatial database. The data cleansing process was employed to exclude unavailable data (such as missing coordinates and missing timestamps). Then, all PU geolocations or taxi zone ID were aggregated into 263 TAZs. Second, we count trips in the same month as monthly ridership for every TAZ.

Figure 4 shows the statistical PU samples for the three types of trips over 36 months. The ridership operated by green taxis was small and decreased slightly over time; ridership of the yellow taxis decreased at a rate of approximately 12% per year. Meanwhile, seasonal variations were also observed, i.e., there were two peak periods from March to May and from September to November in every year; On the contrary, the ridership of TNCs grew very rapidly, especially in July 2017, the monthly ridership of TNC firstly exceeded yellow taxis. Finally, we obtain 9120 valid observations, 348 observations were excluded due to no trip record. According to the literature [18], we carried out the log transformation for the three dependent variables to eliminate the influence of the non-normal distribution.
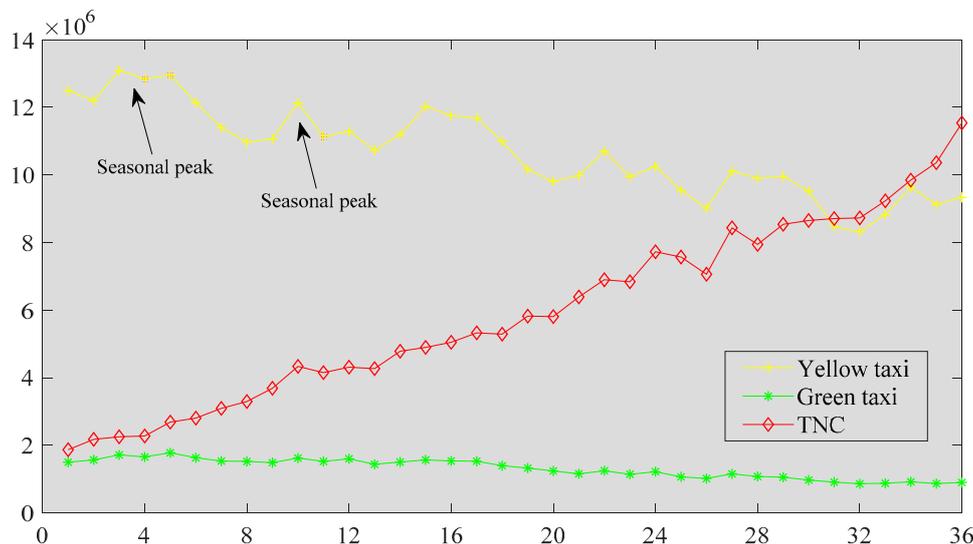

Figure 4. Monthly ridership of three types of taxis in NYC from 2015 to 2017.

4.3. Influencing Factors

Much previous literature has reported that the spatiotemporal distribution of taxis can be affected by a range of external factors. In this case, we extracted four groups of explanatory variables from multiple open datasets, including weather, land use, socioeconomic, and transport condition. Table 2 lists the definitions of all factors, as well as their summary statistics. To be specific, the weather-related variables were downloaded from the NOAA website (https://www.ncdc.noaa.gov), specifically from the NYC station number USW00094728. Four daily ground observations of weather, i.e., snow depth, maximum and minimum temperature, and average wind speed, were selected by calculating their mean value for each month; the second group of land use-related data was downloaded from MapPLUTO®, which is maintained by the NYC Department of City Planning. Considering the case study in [18], we extracted three factors, i.e., residential area, commercial area, and manufacturer area, in each taxi zone. The third group contains five transport-related factors, including the road and bike line lengths and the number of bus stations, subway stations, and bicycle parking zones (called CityRacks). The last group is socioeconomic-related factors, which were obtained from the NYC Geodatabase, which contains eight variables related to demographics, employment, income, vehicle ownership,

education, and commuting. It is important to note that the minimum values for these factors in Table 2 are all zero, because the samples with a default value of zero belong to Central Park, which has taxi ridership data but is missing the corresponding socioeconomic data. The log transformation was also applied for factors from SE1 to SE6 to account for differences in size between TAZs.

Table 2. List of influencing factors.

Group	Label of Factor	Description	Min/Max	Avg
Weather	W1	Number of snowy days in each month	0/7	1.05
	W2	Average maximum temperature in each month (°C)	0.08/30.50	17.78
	W3	Average minimum temperature in each month (°C)	−8.9/22.10	9.74
	W4	Average wind speed in each month (km/h)	1.54/3.34	2.36
Land use	LU1	Percentage of land use for residential purpose in each TAZ (%)	0/96.81	38.40
	LU2	Percentage of land use for commercial purpose in each TAZ (%)	0/64.05	11.93
	LU3	Percentage of land use for manufacturer purpose in each TAZ (%)	0/92.29	9.47
Transport	T1	Length of road per km^2 in each TAZ (/km)	0/58.71	26.22
	T2	Number of subway station per km^2 in each TAZ	0/17.09	1.45
	T3	Number of bus stop per km^2 in each TAZ	0/33	7.21
	T4	Length of bike line per km^2 in each TAZ (/km)	0/16.07	3.55
	T5	Number of CityRacks per km^2 in each TAZ	0/389	38.5
Socioeconomic	SE1	Number of residents with at least Bachelors' degree per km^2 in each TAZ	0/35,295	5723
	SE2	Number of employed residents per km^2 in each TAZ	0/32,885	8474
	SE3	Number of households with more than $75,000 annual income per km^2 in each TAZ	0/18,608	3062
	SE4	Number of vehicle ownership per km^2 in each TAZ	0/2680	1379
	SE5	Number of adults between the ages of 20 and 44 per km^2 in each TAZ	0/22,430	7040
	SE6	Number of employees per km^2 in each TAZ	0/47,037	13,894
	SE7	Average commuting time (minute) in each TAZ	0/60.27	38.83
	SE8	Percentage of commuting to work by public transportation (excluding taxicab) in each TAZ	0/81.92	53.83

5. Model Estimations and Performance

5.1. Selection of Independent Variables

The multicollinearity of the independent variables will cause bias and affect the credibility of the modeling results. To eliminate the collinearity between the factors, we calculate the Pearson correlation coefficient of factors in this study. According to Qian's suggestion [18], if the pairwise correlation coefficients of factors are greater than 0.7, then at most, one of the variables can be included in the model.

Table 3 shows the test results between every two factors. Most of the pairwise correlation coefficients were below 0.7. However, for the weather-related group, all four factors (W1-W4) are highly correlated, so only one of them needs to be retained. Meanwhile, for the socioeconomic factors, the density of residents with at least a Bachelor's degree (SE1) is correlated with the density of employed residents (SE2, 0.92), high income (SE3, 0.99), adults age (SE5, 0.84), and employees (SE6, 0.84), thus these factors (SE2, SE3, SE5 and SE6) need to be removed. Moreover, considering the complex situation of flow at airport, we add a dummy factor to denote whether a TAZ has an airport. In this study, three TAZs containing JFK, EWR, and LGA airport were set to 1, the others were set to 0. As a result, fifteen factors, including thirteen independent variables, the number of month (T), and a dummy variable of airport (AP) are collected and normalized from the initial set of variables.

Table 3. Pearson correlation coefficient for explanatory variables.

Correlations	W1	W2	W3	W4	LU1	LU2	LU3	T1	T2	T3	T4	T5	SE1	SE2	SE3	SE4	SE5	SE6	SE7	SE8
W1	1.00	−0.79	−0.79	0.73	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.00	0.01
W2	−0.79	1.00	0.99	−0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
W3	−0.79	0.99	1.00	−0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
W4	0.73	−0.91	−0.92	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LU1	0.00	0.00	0.00	0.00	1.00	−0.46	−0.46	−0.34	−0.43	−0.16	−0.53	−0.38	−0.32	−0.27	−0.32	0.52	−0.29	−0.21	0.56	−0.08
LU2	0.00	0.00	0.00	0.00	−0.46	1.00	−0.16	0.56	0.67	0.52	0.51	0.52	0.58	0.55	0.59	−0.19	0.54	0.49	−0.57	−0.01
LU3	0.00	0.00	0.00	0.00	−0.46	−0.16	1.00	−0.10	0.01	−0.19	0.02	0.08	−0.10	−0.14	−0.10	−0.40	−0.11	−0.17	−0.16	0.12
T1	0.01	0.00	0.00	0.00	−0.34	0.56	−0.10	1.00	0.47	0.45	0.63	0.38	0.52	0.59	0.52	0.04	0.62	0.57	−0.50	0.28
T2	0.00	0.00	0.00	0.00	−0.43	0.67	0.01	0.47	1.00	0.20	0.46	0.49	0.36	0.36	0.36	−0.23	0.40	0.32	−0.47	0.09
T3	0.00	0.00	0.00	0.00	−0.16	0.52	−0.19	0.45	0.20	1.00	0.43	0.38	0.63	0.70	0.61	0.13	0.69	0.72	−0.37	0.26
T4	0.00	0.00	0.00	0.00	−0.53	0.51	0.02	0.63	0.46	0.43	1.00	0.65	0.61	0.59	0.59	−0.30	0.60	0.55	−0.62	0.24
T5	0.00	0.00	0.00	0.00	−0.38	0.52	0.08	0.38	0.49	0.38	0.65	1.00	0.60	0.56	0.59	−0.22	0.56	0.52	−0.56	0.13
SE1	0.00	0.00	0.00	0.00	−0.32	0.58	−0.10	0.52	0.36	0.63	0.61	0.60	1.00	0.92	0.99	0.04	0.84	0.84	−0.62	0.15
SE2	0.01	0.00	0.00	0.00	−0.27	0.55	−0.14	0.59	0.36	0.70	0.59	0.56	0.92	1.00	0.90	0.22	0.98	0.98	−0.52	0.36
SE3	0.00	0.00	0.00	0.00	−0.32	0.59	−0.10	0.52	0.36	0.61	0.59	0.59	0.99	0.90	1.00	0.05	0.82	0.82	−0.62	0.11
SE4	0.01	0.00	0.00	0.00	0.52	−0.19	−0.40	0.04	−0.23	0.13	−0.30	−0.22	0.04	0.22	0.05	1.00	0.19	0.30	0.41	0.13
SE5	0.01	0.00	0.00	0.00	−0.29	0.54	−0.11	0.62	0.40	0.69	0.60	0.56	0.84	0.98	0.82	0.19	1.00	0.98	−0.51	0.44
SE6	0.01	0.00	0.00	0.00	−0.21	0.49	−0.17	0.57	0.32	0.72	0.55	0.52	0.84	0.98	0.82	0.30	0.98	1.00	−0.44	0.43
SE7	0.00	0.00	0.00	0.00	0.56	−0.57	−0.16	−0.50	−0.47	−0.37	−0.62	−0.56	−0.62	−0.52	−0.62	0.41	−0.51	−0.44	1.00	0.13
SE8	0.01	0.00	0.00	0.00	−0.08	−0.01	0.12	0.28	0.09	0.26	0.24	0.13	0.15	0.36	0.11	0.13	0.44	0.43	0.13	1.00

5.2. Comparison of Model Accuracy

The OLS model is first calibrated to explore significant factors that influence the three dependent variables and the results are presented in Table 4. It shows the estimated coefficients and t-probability for each independent variable and indicators for the goodness-of-fit of the model. Most of the factors are significant at 0.01 level, revealing that these factors are highly related to the ridership for three models. However, several factors are not statistically significant, including W1, T2, and AP for TT model, T2 for TNC model, and LU2 for PoT model. The variance inflation factor (VIF) values of most factors are within a reasonable range (<7.5), indicating that those factors are well selected so that the multicollinearity problem is avoided [33].

Table 4. Estimation results for OLS models.

Variable	TT		TNC		PoT		VIF
	Coefficient	t-prob	Coefficient	t-prob	Coefficient	t-prob	
Intercept	9.499	0.000	6.633	0.000	0.032	0.067	-
W1	−0.090	0.116	−0.248	0.000	−0.047	0.000	1.077
LU1	−0.679	0.000	0.553	0.000	0.179	0.000	2.665
LU2	3.330	0.000	3.549	0.000	0.029	0.050	3.714
LU3	2.955	0.000	3.172	0.000	0.036	0.006	1.971
T1	−1.572	0.000	−0.707	0.000	0.120	0.000	2.610
T2	0.197	0.184	−0.204	0.134	−0.086	0.000	2.271
T3	1.381	0.000	0.528	0.000	−0.150	0.000	2.161
T4	0.598	0.000	0.242	0.047	−0.048	0.001	3.513
T5	1.360	0.000	2.231	0.000	0.142	0.000	2.336
SE1	1.193	0.000	−1.007	0.000	−0.438	0.000	3.258
SE4	3.014	0.000	3.626	0.000	0.123	0.000	2.244
SE7	−12.495	0.000	−7.268	0.000	0.789	0.000	3.838
SE8	6.752	0.000	3.832	0.000	−0.367	0.000	1.630
AP	0.146	0.420	0.440	0.008	0.059	0.004	1.199
T	−0.717	0.000	2.546	0.000	0.467	0.000	1.077
R^2	0.8067		0.6729		0.7333		
R^2_{adj}	0.8064		0.6724		0.7329		
RSS	17692.73		14868.11		221.37		
RMSE	1.2473		1.3680		0.1558		

According to the adjusted R^2, 80.67%, 67.29% of the variation can be explained for the TT and TNC ridership, and 73.33% for the proportion change of TNC. Based on the coefficient values, most of the factors in our study show an intuitive relation with the taxis and TNC ridership, e.g., three factors, including the number of snowy days (W1), length of roads (T1), and commuting time (SE7), are negatively correlated with variation of ridership in both modes. In addition, eight factors, including LU2, LU3, T3, T4, T5, SE4, SE8 and AP show positive effect on increase of TT and TNC ridership. Furthermore, the remaining four factors exhibit different correlations in the two models. For example, the factor of time (T) is negatively correlated with the ridership of TT (−0.721) but positively correlated with the TNC (2.545), which is consistent with the opposite temporal variation presented in Figure 2.

However, for those factors that are homogenous over space and time, it is difficult for the OLS model to explain. For instance, the negative sign of LU1 in TT and TNC ridership models implies that low percentage of residential land use in a TAZ may increase the number of PU points. This situation is contrary to our intuitive understanding. A possible reason is that taxi trips are asymmetric [34] and are more heavily used for trips to residential areas than trips from them. As a result, we conduct further investigations using GTWR models.

The GTWR model needs to estimate each sample independently to obtain coefficients, resulting in voluminous coefficients that vary according to time and place. Table 5 presents the distribution of each

factor for three dependent variables, respectively. The optimal parameter of q is set to 400 and τ is 350 (unit: meter/month) through a CV process via minimization in terms of the R^2. As shown in Table 5, the adjusted R^2 is 0.9787 for TT model and 0.9403 for TNC model and 0.9329 for PoT model, which corresponds to 0.1723 (21%), 0.2679 (39%) and 0.20 (27%) improvement in the amount of variation explained compared to OLS models. Moreover, significant improvements are also achieved for two indicators of residual sum of squares (RSS) and root mean square error (RMSE). It is evident that, by addressing the spatial–temporal heterogeneities effect, the reduction in the RSS and the RMSE values prove the superiority of the GTWR model over the global OLS model in the explanatory power and the goodness of model fit based on the same dataset.

Table 5. Estimation results for GTWR models.

Variable	TT			TNC			PoT		
	LQ	MED	UQ	LQ	MED	UQ	LQ	MED	UQ
Intercept	6.86	12.25	15.81	7.05	10.64	13.80	−0.08	0.17	0.77
W1	−0.14	−0.02	0.08	−0.61	−0.09	0.14	−0.09	−0.01	0.01
LU1	−3.30	−0.66	1.96	−1.67	0.27	2.84	−0.04	0.14	0.38
LU2	−1.79	1.25	6.25	−0.87	1.76	6.44	−0.14	0.08	0.34
LU3	−2.27	1.97	6.43	−1.19	2.22	5.68	−0.12	0.08	0.37
T1	−5.10	−2.54	1.65	−4.23	−1.77	1.18	−0.07	0.11	0.37
T2	−1.08	1.31	6.43	−1.01	0.74	5.50	−0.48	−0.09	0.07
T3	−0.63	0.53	3.60	−0.64	0.55	3.24	−0.20	−0.02	0.11
T4	−4.33	−0.12	2.97	−2.60	0.06	2.41	−0.21	0.01	0.29
T5	−4.63	1.48	8.00	−9.94	0.95	3.61	−1.25	−0.12	0.04
SE1	−6.99	2.09	12.94	−10.30	0.83	6.82	−1.31	−0.19	0.25
SE4	−3.17	0.67	3.15	−2.09	0.51	3.15	−0.17	0.06	0.30
SE7	−16.35	−5.44	0.94	−10.56	−2.20	3.17	−0.16	0.45	1.37
SE8	−2.82	1.07	6.49	−3.73	−0.66	3.36	−0.62	−0.20	0.16
AP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
R^2		0.9787			0.9404			0.9430	
$R^2{}_{adj}$		0.9787			0.9403			0.9329	
RSS		1948.0			2708.1			47.3015	
RMSE		0.4531			0.5259			0.0720	

Moreover, the GTWR model also provides an in-depth understanding of how influencing factors vary locally. The coefficients of the regression model can be used to quantitatively analyze the relationship between influencing factors and the dependent variable. To be specific, if the sign of a coefficient is negative, there is a negative correlation between the factor and dependent variable, which reflects a trend of elimination; otherwise, the factor and dependent variable are positively correlated, indicating a mutually reinforcing relationship. According to the three-column summary, i.e., the lower quartile (LQ), the median (MED), and the upper quartile (UQ), we observed that the median values of the W1, T1, and SE7 are negatively correlated with both TT and TNC ridership, which implies that snowy weather, high-density roads, and lengthy commuting time probably decrease the taxi ridership. It is clear that taxi drivers are less willing to operate on snowy days or traffic congestion caused by high-density roads, resulting in a drop in ridership. Meanwhile, since the lengthy commuting TAZs are mainly located far from the central city, the correlation coefficients are consistent with the actual spatial distribution of taxi/TNC ridership decreasing with the increase of distance from the central zone.

The parameter estimation for the number of the subway station (T2) is always positive in TT and TNC models, which suggests that an increase in subway stations will generate more TT and TNC trips. The positive correlation can be explained in two aspects. First, subway stations are usually crowed thus there is a large passenger volume, which will attract and generate more TT and TNC trips. Secondly, the TT and TNC may be widely used for last-mile trips when passengers get off the subway and commute by TT/TNC to final destinations. Except for these two factors mentioned above, the other

factors show moderate disparity, suggesting that these influencing factors may be positive or negative, which vary significantly over space and time.

6. Discussion

6.1. Temporal Effects of Influencing Factors for TT and TNC Ridership

For the temporal effect of influencing factors on TT and TNC ridership, we take the month as the time interval and use the median of coefficient values of two GTWR models, i.e., TT and TNC to plot the corresponding temporal variation for each influencing factors respectively. According to Figure 5, some interesting findings can be summarized.
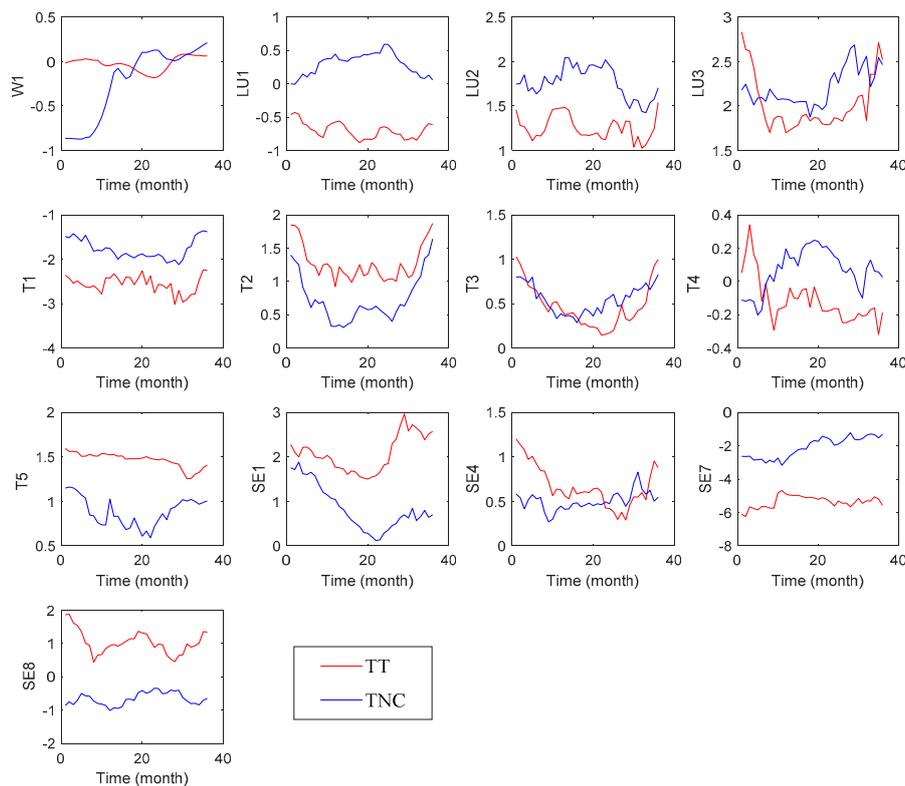

Figure 5. Temporal effects of influencing factor for TT and TNC ridership from 2015 to 2017.

Firstly, the trend of snowy weather (W1) on TT ridership is stable around 0, indicating that ridership of TT is less affected by snowfall weather. Meanwhile, the initial value of TNC is negatively correlated at the beginning of 2015, indicating that snowfall weather will reduce the ridership of TNC, which is mutually verifiable with a previous study [35]. However, the coefficient values of snowy weather on TNC increased dramatically from June 2015 to December 2016 and became competitive with TT. The rapid growth of the coefficient of weather might be contributed to the fact that surge pricing, which was established by TNCs for improving their market competitiveness and quality of service, indeed encourages an increase in supply.

Secondly, for three land use-related factors, the residential land use factor (LU1) shows a negative correlation with taxis but positive with TNC. This pattern is consistent with the OLS model because the independent variable that we chose is PU points rather than DO points. Since the spatial distribution of PU for taxi is asymmetric, i.e., the trips targeting residential areas are larger than those originating from residential areas [34], the coefficients of LU1 for TT are negative. On the contrary, the coefficients of same factor for TNC show a positive correlation because the TNCs serve the outer boroughs more extensively, where residential land use is more prevalent. Another possible reason might because TNC can provide a more personalized service based on the user's current location, rather than relying

on the taxi driver's own experience and habits to pick up passengers. For the commercial land-use factor (LU2) and the manufacturer land-use factor (LU3, mainly refers to the airport, train stations, and external transportation area), the temporal trend of two modes both show significant positive correlation, but the coefficient for TNC is higher than for TT in most of the time. The difference of temporal trends reveals that the increase in the ridership of TNC was more closely related to land use than TT in 2015–2016, resulting in TNC gaining market share rapidly in TAZs with large commercial and manufacturer areas during this period.

Thirdly, for transport-related factors, it can be seen that except for road density, the rest of the Points of Interest (POI) factors in the two models are positively correlated. These temporal variations suggest that taxis and TNC in NYC have mutual promotion effects with other transportation modes, such as buses, subways, and bicycles, reflecting the key role of TT and TNC in meeting the need of the last mile of trips. Moreover, we found that TT is more attractive than TNC where TAZs have more subway stations (T2) and CityRacks (T5), but less attractive at TAZs that have more bus stops and higher densities of bike lines (T4). One possible reason that this pattern occurs is that TT preferred to wait for passengers on POI, while subway stations and CityRacks exist more often near POI. The opposite happens with bus stops, which are spread throughout the city where a TT may be not as available as TNC.

The last group is socioeconomic-related factors, which has an obvious difference between TT and TNC in our case. To be specific, the temporal trend of bachelors' degree factor (SE1) reveals that TT is more attractive to passengers who have higher education, and this has become more obvious in recent years. We assume that this phenomenon is because passenger with higher education might have better chance to make more incomes (0.98 correlate with SE3 in Table 2), and they will use taxis more often; on the other hand, the rapid growth of TNC is observed to be contributed to by a high density of vehicle ownership (SE4) due to the fact that TNC platforms allow people to use an assert (their private car) to make an income [36]. Based on the temporal trend of commuting time (SE7) and public transportation usage rate (SE8), the negative correlation with SE7 infers that for those TAZs that are far away from the city center and have lengthy commuting time, both taxi modes are inadequate to cover the travel needs of these areas, and public transportation might be better choices compared with the expensive cost of taxis and TNC. Meanwhile, the positive coefficients of SE8 factor for the TT further verifies that TTs are most prevalent in central cities, such as Manhattan, where the highly developed public transit network has aggregation effects on TTs.

6.2. Spatial Effects of Influencing Factors for TT and TNC Ridership

Another important advantage of the GTWR model is that the local estimated coefficients that denote local relationships can be mappable and thus allow for visual analysis. It is important to note that, similar to the GWR model, many of GTWR's coefficients might be insignificant, leading to the difficulty to explain heterogeneity in the study area. However, when significance statistics are evaluated and insignificant parameters are removed, the spatiotemporal patterns will become much easier to interpret. In this study, we applied the multiple testing solution proposed by da Silva and Fotheringham [37] to test the significance of local parameter estimates in GTWR to avoid excessive false discoveries. In addition, since the number of local parameter estimates obtained by GTWR at each location corresponds to the valid number of time, it is necessary to assess whether the majority of significant parameters are sufficient to represent the significance of the factor in the TAZ as a whole. In this study, we simply defined an influencing factor in a TAZ as significant when the number of its significant coefficients for all time was greater than 90%. Therefore, we can use the median value of the significant coefficients from GTWR model to produce a spatial variation map for each TAZ.

Taking the coefficients of the PoT model as an example, Figure 6 shows the spatial distribution of coefficients for weather- and land use-influencing factors using graduated colors as rendering style. Figure 6a shows that the spatial distribution of the coefficients for snowy weather is positive in the southern of Manhattan, the central of Staten Island and the JFK airport, which naturally reflects the

fact that TNC trips are aggregated in these TAZs to create higher fares that come from short, frequent trips in midtown or long-distance trips from the airport during snowy days.
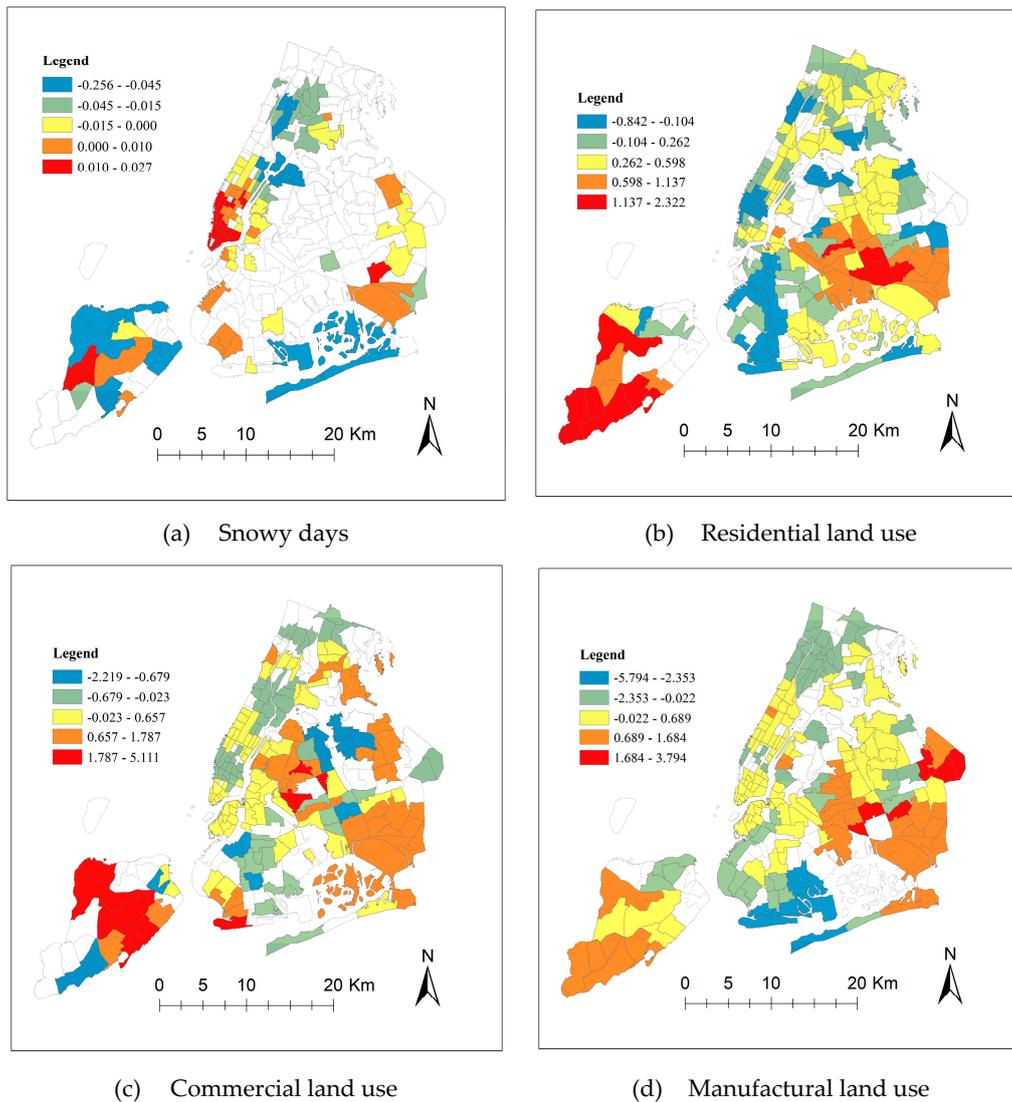

(a)   Snowy days   (b)   Residential land use

(c)   Commercial land use   (d)   Manufacural land use

Figure 6. Spatial distribution for the coefficients of weather and land use factors for PoT.

For the land-use related-factors, Figure 6b–d visualize the spatial distribution of the coefficients of land use for residential, commercial, and manufactural purposes, respectively. In general, the majority of TAZs in Queens and Staten Island are found to be significantly positive for the increment of TNC ridership, while the TAZs in Manhattan, Bronx, and Brooklyn mainly show negative coefficients. Based on the online statistic reports that around 85% of taxi PUs occurred in Manhattan (most of those were made by yellow taxis), it is no surprise that TT has lost its advantages in the outer boroughs where a large number of TNC trips were generated from 2015 to 2017. However, although more land-use area is expected to bring more trips, we found that the land-use patterns are diverse. For example, the residential land-use factor is observed to present more positive effects on increasing TNC ridership in southern Queens, while in eastern Queens, the land use for commercial and manufactural purpose plays a more critical role for the growth of TNC trips. These findings are consistent with the previous analysis reported by Poulsen et al. [38].

The spatial distribution for five transport-related factors is presented in Figure 7a–e. The spatial distribution of road density coefficients in Figure 7a shows that while high-density roads have positive effects on the share of TNC ridership in general (T1, 0.11), such as in Brooklyn and Staten Island,

they also exhibit negative effects in downtown Manhattan and Queens. Normally, TAZs with a higher distribution of other transport POI, such as buses, subways, and bicycles, are correlated with higher passenger density and may produce more taxi/TNC ridership. Figure 7b shows that, in the middle of NYC where the subway system was highly developed, the number of subway stations appear to positively correlation for the increase of proportion for TNC. With respect to the east side of NYC, especially for those areas near the boundary, the negative correlation was observed, revealing this as another pattern of the TNCs experience of growth due to the insufficient of subway stations. Figure 7c shows that the reasons of the increase of TNC trips in Staten Island might be attributed to two aspects: (1) the bus stops in these areas bring in massive amounts of passengers with travel needs; (2) the lack of subway stations and low concentration of TTs in these areas causes these passengers to only rely on TNC, eventually leading to an increase in TNC's market share.

For the socioeconomic-related factors, Figure 8a–d visualize the spatial distribution of the coefficients for SE1, SE4, SE7, and SE8 respectively. The temporal changes of SE1 in Figure 4 imply that TAZs with people who received relatively higher education levels may use TTs more often. On the contrary, areas with more private cars (SE4) may generate more TNC trips. By comparing the Figure 8a,b, we observed a consistent pattern, that the spatial distribution of coefficient for Bachelor's degree is opposite to the distribution of vehicle ownership. To be specific, we found that highly positive coefficients of education are mainly distributed in the remote TAZs, such as East Queens and South Staten Island. As TTs rarely reach these areas and public transportation is inadequate, the probability of hailing a TNC by people with high education/income is significantly increased. Meanwhile, in high-traffic TAZs such as Manhattan, the coefficient is positive due to the better flexibility provided by TNC. It can be seen from Figure 8b that the highest value of coefficients for vehicle ownership was observed at the central of NYC. This indicates that a higher density of private cars in these areas may generate more TNC trips. Although the correlation is positive in general, TAZs in East Queens present negative correlation. A possible explanation is that it is difficult and costly to ride a taxi from these places [18]. As a result, people may use their own vehicles. Finally, the spatial distribution of coefficients for average commuting time and public transportation usage is shown by the following Figure 8c,d. The spatial pattern of the coefficients for commuting time (SE7) takes the greatest around the center of Manhattan, which indicates that the lengthy commuting increases the utility of TNC and thus reducing the taxi ridership. However, considering the fact that public transportation (SE8) is more developed in the west than in the east, the coefficients for SE8 exhibit opposite spatial distribution characteristics as the commuting time. The reasons that SE8 is more positive for TNCs in Western NYC could come from two aspects: the huge amount of passengers brought by the developed public transportation network in the central city, and the absence of TTs in remote areas such as Staten Island.
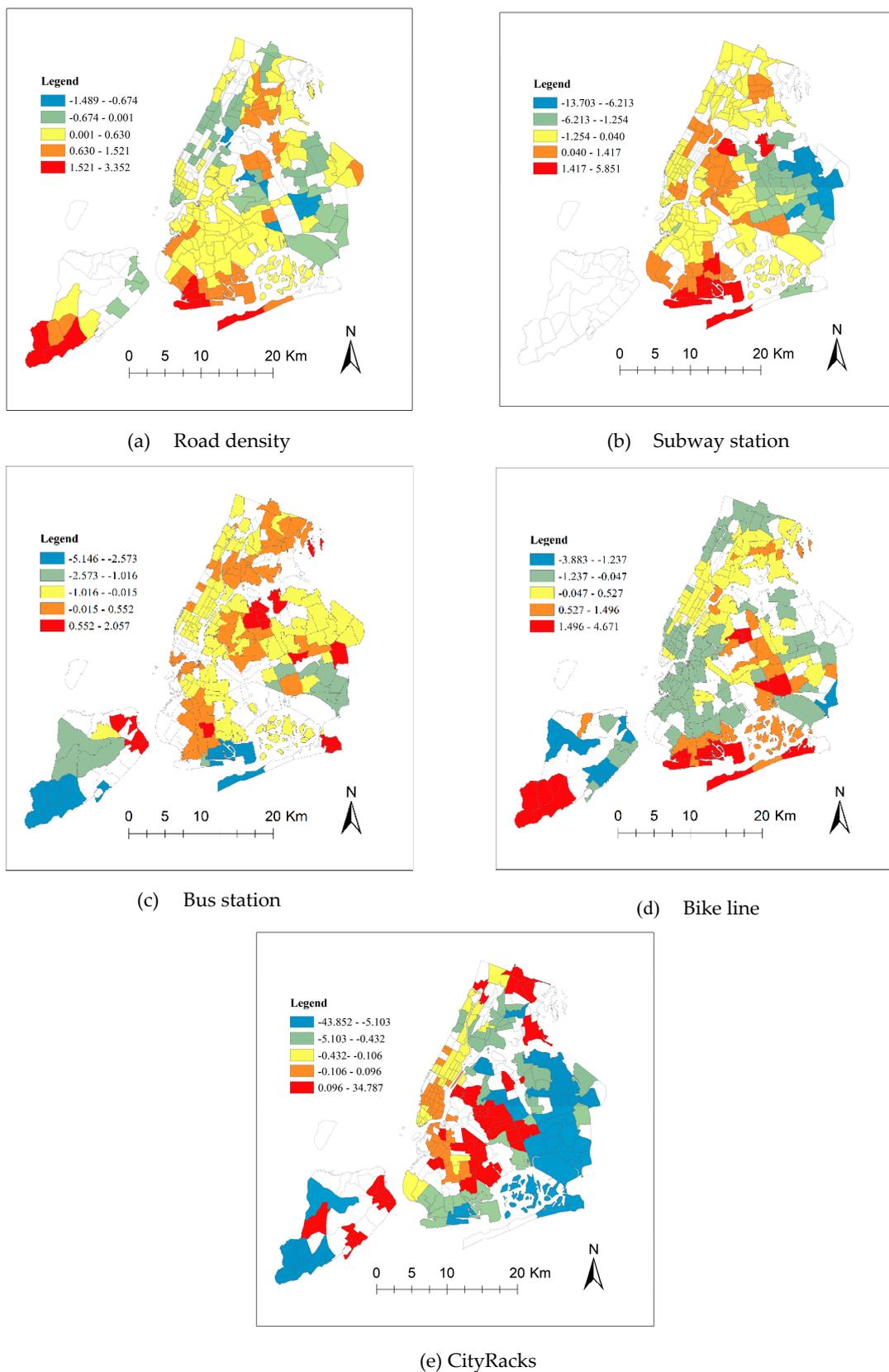
(a)   Road density

(b)   Subway station

(c)   Bus station

(d)   Bike line

(e) CityRacks

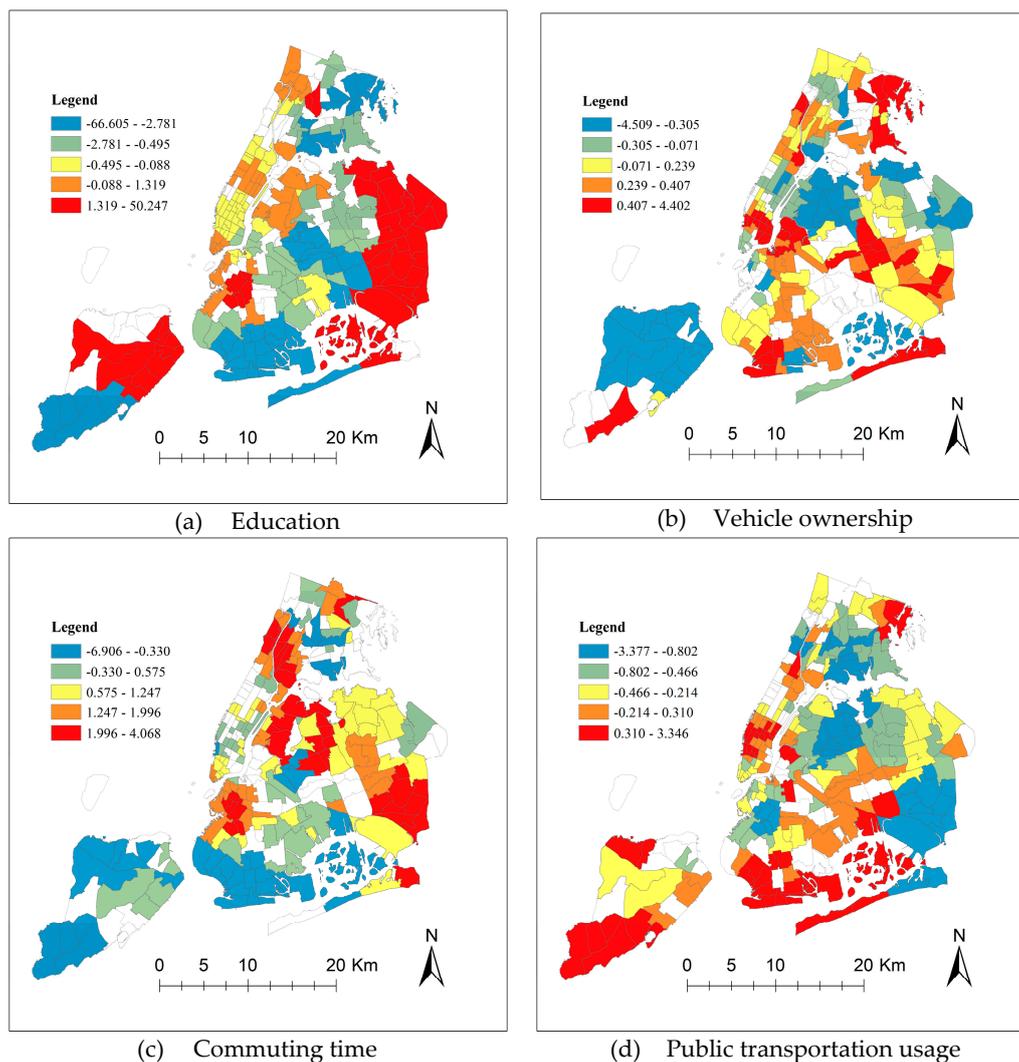**Figure 7.** Spatial distribution for the coefficients of transport factors for PoT.

Figure 8. Spatial distribution for the coefficients of socioeconomic-related factors for PoT.

6.3. The Efficiency of the Parallel-Based GTWR Model

To test the efficiency of the parallel-based GTWR model, we randomly selected the sample set with different proportions and then recoded the calculation time of the basic GTWR model and the improved model. The same process was repeated 10 times to avoid noisy measures due to other processes that could be running at the same time. Table 6 shows a comparison of the average computation time of both models. With the improvement of parallel computing, we observed that the calculation time was reduced by 49% to 61% based on a four-core CPU. Moreover, we found that a 10% sampling level can guarantee the robustness of the algorithm for modeling in our case. However, random sampling is a simple strategy and may cause some valuable information to be omitted, thus weaken the stability of the model result. A better approach is to use systematic or artificial pre-defined methods, based on time-varying features (e.g., cycles and seasonality) to improve the robustness of the algorithm. Moreover, with the popularity of cloud computing technology, the introduction of distributed computing or cloud computing to the GTWR model is also an effective improvement direction [39].

Table 6. Comparison of the efficiency between GTWR and parallel-based GTWR models. (Unit: second).

Percentage of Training Samples	Number of Training Samples	Basic GTWR	Parallel-Based GTWR	Time Reduction
10%	913	13.2	5.7	57%
30%	2739	88.1	34.5	61%
50%	4565	199.6	101.2	49%
100%	9126	1314.4	652.3	50%

7. Conclusions

The rapid development of TNC has been indeed a useful supplement to the traditional taxi industry in the early development stage, but the growth of the urban demand for taxis has been relatively stable. As a result, the relationship between the two modes will inevitably be mutually competitive, and this competitive relationship will demonstrate nonstationarity in time and space. In response to this problem, we select NYC as a case study to illustrate that the GTWR model can be an effective tool for analyzing spatiotemporal heterogeneity. Moreover, the effects of the influencing factors for the TT and TNC can be quantitatively evaluated in the temporal term, and spatial variations can also be analyzed by the coefficients at different spatial units (i.e., administrative division-based or grid-based).

This study compares GTWR with OLS while exploring the relationships between built environment and the PU ridership. The global coefficients of OLS models are observed to be deficient when dealing with spatial problem. The GTWR model, on the other hand, shows better performance than the OLS model, especially in the fact that the GTWR model can help to eliminate potential bias from spatiotemporal heterogeneity and provide localized regression statistics at each location. By visualizing distributions of median values of coefficients for each factor, the spatiotemporal variations of the factors could be better interpreted. Our study demonstrates that the relationships between ridership and influencing factor of built environment vary over space and time in NYC. Moreover, the effects of influencing factors on TT and TNC are significantly different on both spatial and temporal terms. For example, the model results reveal that the TNC's surge pricing policy has a significant effect on increasing TNC trips in snowy conditions, especially in western Manhattan. While TTs have always been dominant in downtown Manhattan, the share of TNC has risen significantly in the adjacent neighborhoods due to the availability of transit alternatives, such as subways, buses, and private cars, which is probably correlated with commuting time (SE7). Meanwhile, the increases of TNCs are also observed in remote places, which are positively correlated with densities of multiple land use, educated populations, and levels of public transportation usage. Compared to the current saturation of demand in the central city, future competition between TT and TNC might be concentrated in remote areas, such as eastern Queens, which is not adequately covered by public transportation. We believe these findings of spatial variations of taxi demand could provide useful scientific guidelines for the taxi industry and TNC to optimize their existing resources thus improving efficiency. Furthermore, the basic modeling steps described in this paper, such as data aggregation, factor selection, parameter optimization, modeling analysis, and visual presentation, can also be applied to other research fields for spatiotemporal modelling. For example, considering the recent outbreak of Coronavirus Disease 2019 (COVID-19), the GTWR model might be an appropriate approach to assess the local relationships between the contagiousness of the virus and the influencing factors of urban environment.

Several challenges remain when applying GTWR models to explore detailed variations in relationships between taxis and built environment research. As the variation in transportation environments in different cities is enormous, the result of GTWR can only be adapted to specific cities. In the follow-up study, we will apply the GTWR model to other large cities for comparison and evaluation. The model will incorporate with different type of influencing factors, such as POI, real-time population flow, and Internet of Things data, which will help to improve the interpretation of how the urban spaces and times result in taxi demand.

We also notice that using a four-core CPU might be insufficient to fully evaluate the performance of the proposed parallel-based GTWR model. In fact, the optimization of computational performance for GWR-based models is always a technical bottleneck that plagues the widespread application of spatiotemporal modeling, especially in the face of a massive spatial and temporal dataset. In this study, the most important thing that we focus on is to apply the GTWR model to evaluate the relationship between the taxi ridership and the influencing factors of built environment. Due to the length limitation of this paper, we only provide a simple technical idea of the design of the parallel-based GTWR model and perform it with a small-scale case study (less than 10,000 samples). The efficiency of the current use of a four-core CPU already meets the needs of this study. According to some literature [39], the efficiency of parallel computing is correlated with many factors, such as the structure of the algorithm, the selection of software, and the size of the data, and in some cases the computation is even less efficient than serial computation. Therefore, evaluating the efficiency of the parallel-based GTWR model could be a complex technical problem, which we believe is necessary to conduct in future.

Another issue that must be considered is that the GTWR model uses statistical local least squares to estimate the coefficient of variables; therefore, the model's accuracy depends on the independence of the observed samples. When there is a strong autocorrelation between the data samples, and this autocorrelation is not considered well, it will cause the overfitting problem and affect the final explanatory results of the model. In this respect, the GTWAR model that can estimate the spatial autocorrelation for each variable might be a better solution. However, the GTWAR model increases the computational complexity of the algorithm; therefore, whether this model is necessary for transportation analysis needs to be further evaluated. Regarding the time dimension, seasonal change might be considered. There are more taxis to be operated in summer than in winter, although, in our study, we use weather factors to reflect seasonal changes. The GcTWR model proposed in [26] might be an alternative way to improve the general GTWR model. The question is that the definition of the seasonal span of the GcTWR model is manually preset. In actual situations, periodicity varies. Adopting a certain adaptive method to auto-identify the seasonal span of transportation distribution is another problem that must be considered carefully.

Author Contributions: Conceptualization, Xinxin Zhang and Bo Huang; methodology, Bo Huang; software, Bo Huang; validation, Xinxin Zhang, Bo Huang, and Shunzhi Zhu; data curation, Shunzhi Zhu; writing—original draft preparation, Xinxin Zhang; writing—review and editing, Xinxin Zhang and Bo Huang; visualization, Xinxin Zhang; supervision, Bo Huang; and funding acquisition, Shunzhi Zhu. All authors have read and agreed to the published version of the manuscript.

References

1. Wallsten, S. The competitive effects of the sharing economy: How is uber changing taxis. *Technol. Policy Inst.* **2015**, *22*, 1–22.
2. Rayle, L.; Shaheen, S.; Chan, N.; Dai, D.; Cervero, R. *App-Based, on-demand Ride Services: Comparing Taxi and Ridesourcing Trips and User Characteristics in San Francisco University Of California Transportation Center (UCTC)*; University of California: Berkeley, CA, USA, 2014.
3. Nie, Y.M. How can the taxi industry survive the tide of ridesourcing? Evidence from shenzhen, China. *Transp. Res. Part C Emerg. Technol.* **2017**, *79*, 242–256. [CrossRef]
4. Wang, X.; He, F.; Yang, H.; Gao, H.O. Pricing strategies for a taxi-hailing platform. *Transp. Res. Part E Logist. Transp. Rev.* **2016**, *93*, 212–231. [CrossRef]

5. Cramer, J.; Krueger, A.B. Disruptive change in the taxi business: The case of uber. *Am. Econ. Rev.* **2016**, *106*, 177–182. [CrossRef]

6. Hall, J.V.; Krueger, A.B. An analysis of the labor market for uber's driver-partners in the united states. *ILR Rev.* **2018**, *71*, 705–732. [CrossRef]

7. Li, M.; Dong, L.; Shen, Z.; Lang, W.; Ye, X. Examining the interaction of taxi and subway ridership for sustainable urbanization. *Sustainability* **2017**, *9*, 242. [CrossRef]

8. Chen, C.; Varley, D.; Chen, J. What affects transit ridership? A dynamic analysis involving multiple factors, lags and asymmetric behaviour. *Urban Stud.* **2011**, *48*, 1893–1908. [CrossRef]

9. Yang, Z.; Franz, M.L.; Zhu, S.; Mahmoudi, J.; Nasri, A.; Zhang, L. Analysis of washington, dc taxi demand using gps and land-use data. *J. Transp. Geogr.* **2018**, *66*, 35–44. [CrossRef]

10. Castro, P.S.; Zhang, D.; Chen, C.; Li, S.; Pan, G. From taxi gps traces to social and community dynamics: A survey. *ACM Comput. Surv. (CSUR)* **2013**, *46*, 17. [CrossRef]

11. Chiou, Y.-C.; Jou, R.-C.; Yang, C.-H. Factors affecting public transportation usage rate: Geographically weighted regression. *Transp. Res. Part A Policy Pract.* **2015**, *78*, 161–177. [CrossRef]

12. Liu, Y.; Wang, F.; Xiao, Y.; Gao, S. Urban land uses and traffic 'source-sink areas': Evidence from gps-enabled taxi data in shanghai. *Landsc. Urban Plan.* **2012**, *106*, 73–87. [CrossRef]

13. Sadowsky, N.; Nelson, E. The impact of ride-hailing services on public transportation use: A discontinuity regression analysis. 2017. Economics Department Working Paper Series. 13. Available online: https://digitalcommons.bowdoin.edu/econpapers/13 (accessed on 26 May 2017).

14. Hochmair, H.H. Spatiotemporal pattern analysis of taxi trips in new york city. *Transp. Res. Rec. J. Transp. Res. Board* **2016**, *2542*, 45–56. [CrossRef]

15. Fotheringham, A.S.; Crespo, R.; Yao, J. Geographical and temporal weighted regression (gtwr). *Geogr. Anal.* **2015**, *47*, 431–452. [CrossRef]

16. Chow, L.-F.; Zhao, F.; Liu, X.; Li, M.-T.; Ubaka, I. Transit ridership model based on geographically weighted regression. *Transp. Res. Rec.* **2006**, *1972*, 105–114. [CrossRef]

17. Cardozo, O.D.; García-Palomares, J.C.; Gutiérrez, J. Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Appl. Geogr.* **2012**, *34*, 548–558. [CrossRef]

18. Qian, X.; Ukkusuri, S.V. Exploring spatial variation of urban taxi ridership using geographically weighted regression. In Proceedings of the Transportation Research Board 94th Annual Meeting, Washington, DC, USA, 11–15 January 2015.

19. Fotheringham, A.S.; Crespo, R.; Yao, J. Exploring, modelling and predicting spatiotemporal variations in house prices. *Ann. Reg. Sci.* **2015**, *54*, 417–436. [CrossRef]

20. Wu, J.; Yao, F.; Li, W.; Si, M. Viirs-based remote sensing estimation of ground-level pm2. 5 concentrations in beijing–tianjin–hebei: A spatiotemporal statistical model. *Remote Sens. Environ.* **2016**, *184*, 316–328. [CrossRef]

21. Taylor, B.D.; Miller, D.; Iseki, H.; Fink, C. *Analyzing the Determinants of Transit Ridership Using a Two-Stage Least Squares Regression on a National Sample of Urbanized Areas*; University of California Transportation Center: Berkeley, CA, USA, 2003.

22. Huang, B.; Wu, B.; Barry, M. Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 383–401. [CrossRef]

23. He, Q.; Huang, B. Satellite-based mapping of daily high-resolution ground pm 2.5 in china via space-time regression modeling. *Remote Sens. Environ.* **2018**, *206*, 72–83. [CrossRef]

24. Chu, H.-J.; Kong, S.-J.; Chang, C.-H. Spatio-temporal water quality mapping from satellite images using geographically and temporally weighted regression. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *65*, 1–11. [CrossRef]

25. Wu, B.; Li, R.; Huang, B. A geographically and temporally weighted autoregressive model with application to housing prices. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 1186–1204. [CrossRef]

26. Du, Z.; Wu, S.; Zhang, F.; Liu, R.; Zhou, Y. Extending geographically and temporally weighted regression to account for both spatiotemporal heterogeneity and seasonal variations in coastal seas. *Ecol. Inform.* **2018**, *43*, 185–199. [CrossRef]

27. Ma, X.; Zhang, J.; Ding, C.; Wang, Y. A geographically and temporally weighted regression model to explore the spatiotemporal influence of built environment on transit ridership. *Comput. Environ. Urban Syst.* **2018**, *70*, 113–124. [CrossRef]

28. Zhang, X.; Huang, B.; Zhu, S. Spatiotemporal influence of urban environment on taxi ridership using geographically and temporally weighted regression. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 23. [CrossRef]

29. Liu, Y.; Ji, Y.; Shi, Z.; Gao, L. The influence of the built environment on school children's metro ridership: An exploration using geographically weighted poisson regression models. *Sustainability* **2018**, *10*, 4684. [CrossRef]

30. Peruggia, M. Model selection and multimodel inference: A practical information-theoretic approach (2nd ed.).(telegraphic reviews)(book review). *J. Wildl. Manag.* **2002**, *67*, 175–196.

31. Sharma, G.; Martin, J. Matlab®: A language for parallel computing. *Int. J. Parallel Program.* **2009**, *37*, 3–36. [CrossRef]

32. Yazici, M.A.; Kamga, C.; Singhal, A. A big data driven model for taxi drivers' airport pick-up decisions in New York city. In Proceedings of the 2013 IEEE International Conference on Big Data, Silicon Valley, CA, USA, 6–9 October 2013; pp. 37–44.

33. Neter, J.; Kutner, M.H.; Nachtsheim, C.J.; Wasserman, W. *Applied Linear Statistical Models*; McGraw-Hill Irwin: New York, NY, USA, 1996.

34. King, D.A.; Peters, J.R.; Daus, M.W. Taxicabs for improved urban mobility: Are we missing an opportunity? In Proceedings of the Transportation Research Board 91st Annual Meeting, Washington, DC, USA, 22–26 January 2012.

35. Kamga, C.; Yazici, M.A.; Singhal, A. Hailing in the rain: Temporal and weather-related variations in taxi ridership and taxi demand-supply equilibrium. In Proceedings of the Transportation Research Board 92nd Annual Meeting, Washington, DC, USA, 13–17 January 2013.

36. Davidov, G. The status of uber drivers: A purposive approach. *Span. Labour Law Employ. Relat. J. Forthcom. (2017)* **2016**. Forthcoming. [CrossRef]

37. Da Silva, A.R.; Fotheringham, A.S. The multiple testing issue in geographically weighted regression. *Geogr. Anal.* **2016**, *48*, 233–247. [CrossRef]

38. Poulsen, L.K.; Dekkers, D.; Wagenaar, N.; Snijders, W.; Lewinsky, B.; Mukkamala, R.R.; Vatrapu, R. Green cabs vs. Uber in New York city. In Proceedings of the 2016 IEEE International Congress on Big Data (BigData Congress), San Francisco, CA, USA, 27 June–2 July 2016; pp. 222–229.

39. Li, Z.; Fotheringham, A.S.; Li, W.; Oshan, T. Fast geographically weighted regression (fastgwr): A scalable algorithm to investigate spatial process heterogeneity in millions of observations. *Int. J. Geogr. Inf. Sci.* **2019**, *33*, 155–175. [CrossRef]