Transport Policy 115 (2022) 75–87

Contents lists available at ScienceDirect

Transport Policy

journal homepage: www.elsevier.com/locate/tranpol

Evaluating rail transit's comparative advantages in travel cost and time over taxi with open data in two U.S. cities

Sajeeb Kirtonia, Yanshuo Sun *

Department of Industrial and Manufacturing Engineering, FAMU-FSU College of Engineering, Florida State University, 2525 Pottsdamer Street, Tallahassee, FL, 32310, USA

ARTICLE INFO

Keywords:
Rail transit
Comparative analysis
Taxi trip data
Public transport marketing

ABSTRACT

This paper presents a comparative analysis of rail transit and taxi by travel cost and time based on the large-scale taxi trip data and public transit schedule information in two major U.S. cities. To quantify the relative advantage of one mode over the other, we introduce the notion of travel gradient, which is travel cost difference divided by travel time difference. Based on the signs of travel cost and time differences, we classify all trips into four quadrants. Quadrant II trips are selected for further analysis because rail transit is identified to be competitive with taxi for such trips. We also explore the relation between various trip characteristics and travel gradient with and without considering the spatial variation of such a relation. Main research findings include: (1) around 70% of the taxi trips in the considered datasets can be substituted with rail transit trips if the maximum walking distance is 0.5 miles at each trip end; (2) for around 10% of taxi trips with both modes being viable, rail transit dominates taxi in both travel cost and time; for the rest, rail transit is competitive with taxi; (3) the marginal travel cost saving due to mode switching from taxi to rail transit is about $70; and (4) there exist clearly spatial variations of the relation between trip characteristics and travel gradient. The main policy recommendation from this study is that rail transit can be better marketed by highlighting its relative advantage over taxi in travel time and cost, especially for travels in certain directions and time periods.

1. Introduction

The objective of this study is to empirically evaluate the advantages of traveling by rail transit in travel cost and travel time over taxi travels, based on the large-scale taxi trip data and public transit schedule information in Washington, D.C. and Chicago, IL. Taxi and rail transit are two important, while distinct, transportation modes in the urban environment. Due to the high flexibility, taxi is desirable for travelers who require personal, door-to-door, and on-demand mobility services (Sun and Zhang, 2018). Nonetheless, taxi trips tend to be costly and subject to traffic congestion. Rail transit is characterized by its dedicated right-of-way, fixed routes, and fixed schedules. Thus, rail transit travels are usually more reliable than other transportation modes in the urban area (Sun and Xu, 2012). Rail transit is also more affordable as rail transit operations are usually subsidized, especially in the U.S. An increasing number of U.S. cities, such as New York City (NYC), Chicago, and Washington, D.C., make available their taxi trip data through open data initiatives, which provides an unprecedented opportunity to investigate the taxi performance. Rail transit schedules and fares are also easily accessible, as rail transit operators have developed web-based and mobile trip planners. Therefore, based on the publicly available data in D.C. and Chicago, this paper seeks to present an empirical study to compare rail transit and taxi by travel time and cost, and also explore how various trip characteristics relate to one mode's competitiveness over the other.

A review of the relevant literature shows that many studies have characterized taxi trip characteristics and compared taxi with transit by trip patterns (spatial and temporal distributions of trips), such as Hochmair (2016) and Wang and Ross (2019). In particular, the NYC taxi trip data are frequently used in the literature mainly because of its widely known availability. While several studies have compared non-driving transportation modes, such as taxi vs bike-sharing, there are no systematic comparisons of rail transit and taxi by travel cost and travel time based on large-scale empirical data. Thus, it is also unclear how the comparative advantage is related to various trip characteristics, especially when spatial variations of this relation are considered. This paper thus seeks to fill those gaps by presenting a comparative study of rail transit and taxi by travel cost and time with the real-world data from

* Corresponding author.
E-mail addresses: sk18da@my.fsu.edu (S. Kirtonia), y.sun@eng.famu.fsu.edu (Y. Sun).

https://doi.org/10.1016/j.tranpol.2021.11.003
Received 3 May 2021; Received in revised form 28 September 2021; Accepted 6 November 2021
Available online 10 November 2021

Washington, D.C. and Chicago. The focus of this paper is on the comparison of taxi and rail transit; multimodal transit involving transfers between rail transit and other modes, such as bus, is not considered in this study.

We first collect and process taxi trip data and rail transit travel information. Then, the notion of travel gradient, based on the relation between the travel cost difference and the travel time difference, is proposed to quantify one mode's comparative advantage over the other. A multiple linear regression model is used to explore how various trip characteristics relate to travel gradient. Then, we examine the spatial variation of the relations between the trip characteristics and travel gradient using geographically weighted regression. While there are a few important findings from our empirical analyses, the main takeaway is that due to the high competitiveness of rail transit in the urban environment, targeted rail transit marketing can be launched in certain urban areas to increase rail transit ridership by highlighting the relative advantage of rail transit over taxi.

The remainder of this paper is structured as follows. After the review of relevant studies in Section 2, we describe how data are collected and cleaned in Section 3. Then, a metric is designed and interpreted in Section 4, which is used as the response variable in the regression analyses. Case studies are conducted and results are interpreted in Sections 5, followed by discussions in Sections 6. The last section presents a brief summary and identifies prospective research directions.

2. Literature review

2.1. Taxi trip pattern analysis and its comparison with public transit

Many studies have analyzed taxi trip patterns (e.g., spatial distribution of trip origins and destinations, temporal distribution of trips) based on taxi trip or trajectory data. We review only a few example studies, because a comparison is not involved in such taxi-only studies. With taxi trajectory data collected in Shanghai, China, Liu et al. (2012) analyzed the temporal and spatial distributions of pickup and drop-off locations, distribution of trip directions, and distribution of trip lengths. By noticing the close relations between intra-city travel patterns and city structures, Liu et al. (2015) used the community detection method (a method to partition a network into closely connected sub-networks) to identify the sub-regional city structure based on the taxi trip data in Shanghai. They then analyzed the hierarchical and polycentric structure of Shanghai. Hochmair (2016) explored basic taxi trip characteristics in NYC, such as temporal distribution of taxi trips, trip distance distribution, and temporal variations of travel speed. A negative binomial regression model was also presented to explore the relation between taxi trips and other explanatory variables, such as population and employment data, socioeconomic factors, built environmental variables and presence of airports.

We continue to review some studies on the comparison of trip patterns of taxi and public transit. Kim (2018) compared the trip patterns of subway and taxi in Seoul, Korea. They found the number of subway trips was ten times larger than taxi trips. The temporal distributions of trips were quite different for two modes: subway had two distinct travel peaks in the morning and afternoon, respectively, while no sharp travel peaks were observed for taxis. The influence of various explanatory variables on the trip pattern was also explored through classification. Using Singapore as a case study, Zhang et al. (2018) conducted a comparative study of taxi and public transit based on (1) the spatial distributions of trips, (2) the distance decay of travels, and (3) the spatial interactions of urban spaces. They found the spatial distributions of taxi and public transit trips were highly correlated; the public transit travel distance tended to decay faster than taxi trips; travels by two modes also revealed the polycentric urban structure of Singapore.

Based on the spatial relations between taxi trip origins/destinations and subway stations, Wang and Ross (2019) classified all taxi trips in NYC into three categories, namely transit-competing, transit-complementing, and transit-extending. For instance, for transit-extending taxi trips, taxis provide access to or egress from train stations; for transit-complementing trips, taxis serve passengers in areas and during times where transit is unavailable. Trips that could be replaced by taking transit are defined as transit-competing trips. The authors found that a substantial portion of taxi trips (58.53% of 1 million trips) were transit-competing. The authors also tried to explore the demographic characteristics of taxi riders and found that around 60% of taxi trips served economically and physically disadvantaged individuals. Jiang et al. (2018) adopted a similar analysis framework and conducted a case study using data from Beijing, China. Ma et al. (2015) studied a similar topic by exploring whether bike-sharing complemented rail transit or substituted it with data in Washington, D.C. Their regression analyses confirmed a positive correlation between transit ridership and bike-sharing demand at the station level and concluded that a 10% increase in bike-sharing demand would generate a 2.8% increase in rail transit demand. Irawan et al. (2020) followed this line of research and compared motorcycle-based ridesourcing, motorcycle taxi and public transit with survey data from the Jakarta metropolitan area in Indonesia.

2.2. Comparison of nondriving modes by cost and time

We then review empirical studies comparing nondriving transportation modes (e.g., public transit, taxi, active modes) by travel time and/or cost in the urban environment. Other modes such as bike-sharing are included in the review because there are very few empirical studies that directly compared taxi and rail transit.

Faghih-Imani et al. (2017) compared two urban travel modes, taxi and bike-sharing, based on the empirical data on travel time in NYC in 2014. Taxi trips with origins and destinations located in the service area of CitiBike (a bicycle sharing service provider in NYC) were selected for comparison. Travel times were compared by time of day and day of week. They concluded that on average taxi trips were slightly faster than bike-sharing trips in dense urban areas. They also used a logit model to better understand the effect of various factors on the competitiveness of those two modes. In Faghih-Imani et al. (2017), travel time is the only criterion with no consideration of travel cost.

Yang et al. (2014) focused on the comparison of subway and taxi, which were two competing airport ground access modes. They built a binary logit model to analyze travels between Pennsylvania Station in NYC and three major airports serving NYC, namely John F. Kennedy International Airport, Newark Liberty International Airport, and LaGuardia Airport. Their results showed that transit dominated taxi in the airport ground access market for most of the time except during the midnight. They also examined the impact of group size and value of time on the choice between two modes. Although both travel cost and time were involved, Yang et al. (2014) focused on a special market, namely airport ground access. Therefore, specific findings from Yang et al. (2014) may not hold on a larger scale, such as on the city level.

Li et al. (2018) studied the mode choice between taxi and rail transit using taxi trip data and travel survey data. They found convenience, which was quantified as a mixture of travel distance (especially access walking) and travel time, was the dominant factor in influencing which mode to choose between taxi and rail transit. Travel cost was not explicitly modelled, because during the study period in 2014, a flat fare of 2 Chinese Yuan (approximately 0.3 U.S. dollars) was used for Beijing's rail transit travels regardless of the travel distance.

Ulak et al. (2020) used the NYC taxi trip data to explain why taxi was a major transportation mode despite its high cost and concluded that convenience (encompassing easy access, high comfort level, etc.) was the major advantage of taxi. To estimate the value of convenience, they compared taxi trips with so-called equivalent rail transit trips. A rail transit trip was considered equivalent, if (1) a taxi trip origin and destination were within 200 m of a rail transit station and (2) the rail transit travel did not involve a transfer. Clearly, this comparison was

limited to a small fraction of urban travels, due to the strict limits on the access and egress distances (i.e., 200 m) and the number of transfers (i.e., no transfers).

Comparing transportation modes by travel time and cost is important, because both travel time and travel cost are main determinants of mode choice (Pinjari and Bhat, 2006), although other factors, such as comfort and convenience, may also play a role. With household travel survey data in Seoul, Korea, Ha et al. (2020) found that travelers across multiple age and income categories were significantly affected by the travel time and cost differences in their mode choices. It was also indicated that systematic comparisons of travel time and cost among mode alternatives was essential to the formulation of effective policy interventions for shifting travelers' choices toward transit.

2.3. Summary

Due to the increasing data availability on taxi travels, many studies have characterized taxi trips and explored how taxi demand is related to other built environment and socioeconomic variables. NYC frequently appears in such taxi studies, among other U.S. cities, partially because of its widely known availability of taxi trip data. The travel pattern difference between transit and taxi has also been well explored; however, there are no studies that have systematically compared rail transit and taxi by travel cost and time. In addition, little is known about the spatial variation of the comparative advantage of rail transit over taxi. Therefore, this study aims to fill the above research gaps.

3. Data

Two major cities in the U.S., namely Washington, D.C. and Chicago, IL, were selected for the following reasons: (1) both cities made data on taxi trips publicly available; (2) online rail transit trip planners were provided by the rail transit operators; and (3) the two rail transit systems (Washington Metro and Chicago "L") were comparable by the annual ridership, system length, and number of rail lines. By ridership, Washington Metro ranked the second among all rail transit systems in the U.S. as of 2019, and Chicago "L" ranked the third; by the number of stations, Chicago "L" was in the second place, and Washington Metro was the third. The NYC Subway was not included in this analysis, because the NYC Subway was significantly larger than all other U.S. rail transit systems in every aspect. For example, as of 2019 the average weekday ridership of the NYC Subway was about ten times as large as the ridership of Washington Metro (Wikipedia contributors, 2019). All the data described in this study were collected in June 2019.

3.1. Taxi trip data

The taxi trip data in Washington, D.C., and Chicago were freely available on opendata.dc.gov and data. cityofchicago.org, respectively. Due to the data availability issue (the 2018 taxi trip data were not available online for Washington, D.C. as of June 2019), we selected the first seven days in June 2017 (i.e., June 1, 2017–June 7, 2017) as the study period. Therefore, only taxi trips with timestamps in this period were used in this study.

While the original datasets contained other fields, we selected the following: trip origin (latitude and longitude), trip destination (latitude and longitude), trip start time, trip end time, trip duration, mileage, and total fare (including toll, surcharge, and gratuity). To avoid privacy issues, several measures were implemented by the data providers (e.g., the City of Chicago). For example, in the Chicago dataset trip start and end times were rounded to the nearest 15 min, and trip origins/destinations were replaced by the centroids of Census Tracks and Community Areas (City of Chicago, 2016). In Washington, D.C., trip locations and timestamps seemed original, although similar data processing was mentioned in the metadata file, such as rounding pickup and drop-off times to the nearest hour.

As the Chicago dataset contained taxi trips with both ends located within the city limits, for consistency we removed those taxi trips with origins or destinations outside the boundary of Washington, D.C. To clean the taxi trip datasets, for both datasets we removed those trips that had missing values in the selected columns, the same origin and destination, erroneous trip duration (e.g., less than 2 min or more than 60 min), erroneous trip mileage (e.g., less than 1 mile or more than 40 miles).

After data explorations, we noticed a few major discrepancies and inconsistencies in the D.C. taxi dataset. We adopted the following data quality control measures. First, regarding travel time we found the units of trip duration (minutes vs seconds) were inconsistent. We recalculated taxi trip duration using the original trip start and end times for D.C. taxi trips. Regarding travel distance, the haversine function was used to estimate the trip distance based on the trip origin and destination to benchmark the original mileage. As the haversine function returns the great-circle distance, which is the shortest distance between two locations on earth, we further removed those trips with mileage smaller than the great-circle distance or three times larger than it. After filtration, we computed the average travel speed as the total mileage divided by the trip duration. We then removed trips with an average travel speed larger than 60 miles per hour, because an average travel speed of 60 miles seems a good indicator of erroneous trip attributes, time or distance.

Fig. 1 shows the average taxi travel speed by time period of day and day of week for taxi trips after data cleaning. For both cities, the travel speed is relatively high in early morning and it drops significantly in morning peak hours (8–10 a.m.) and afternoon peaks (4–7 p.m.) on workdays. During the weekends, the average travel speed is steady over time, especially in Washington, D.C. Overall, the taxi travel speed in Chicago is higher than the speed in D.C.

3.2. Rail transit travel information

Empirical data on rail transit travel cost and time in both cities (such as the automatic fare collection data Sun and Schonfeld (2016)) were not publicly available. Therefore, we used the schedule and fare information provided by the two rail transit operators through their web-based trip planners. As generally rail transit travels are free of traffic congestion and other incidents, the rail transit travel time should be quite reliable: the scheduled travel time is very close to the actual travel time (Sun and Xu, 2012).

We obtained rail transit travel information in D.C. from the trip planner by the WMATA (Washington Metropolitan Area Transit Authority), which was available at www.wmata.com. A Python script was developed to automate the query. In each request, we specified the origin, destination, departure time, travel mode (rail only), and walking distance limit; in each response, we recorded the access walking time, transit time, egress walking time, access distance, egress distance, total trip time, regular fare, and number of transit transfers. In case of multiple travel plans, only the first plan was kept. We found that for some unknown reason the WMATA trip planner could not return valid rail transit travel plans for about 10% of the travel requests, raising the error "no service at origin/destination at the date/time specified." After thorough checking, we confirmed valid travel plans did exist for such requests. We therefore avoided such an issue by slightly shifting the departure time (e.g., by 2 min) in case that particular error code was returned by the trip planner. Note that in D.C. the rail transit fare varies with time of day and travel distance.

For Chicago "L" operated by the Chicago Transit Authority (CTA), we used the trip planner at www.transitchicago.com/planatrip/. There were two available trip planning engines, Google Maps (GM) and RTA (Regional Transportation Authority) Trip Planner. The RTA trip planner accepted street addresses by default, while in our analysis we used geographic coordinates of locations. To avoid complications in the conversion of geographic coordinates to street addresses, we chose GM because geographic coordinates could be accepted by GM, thus

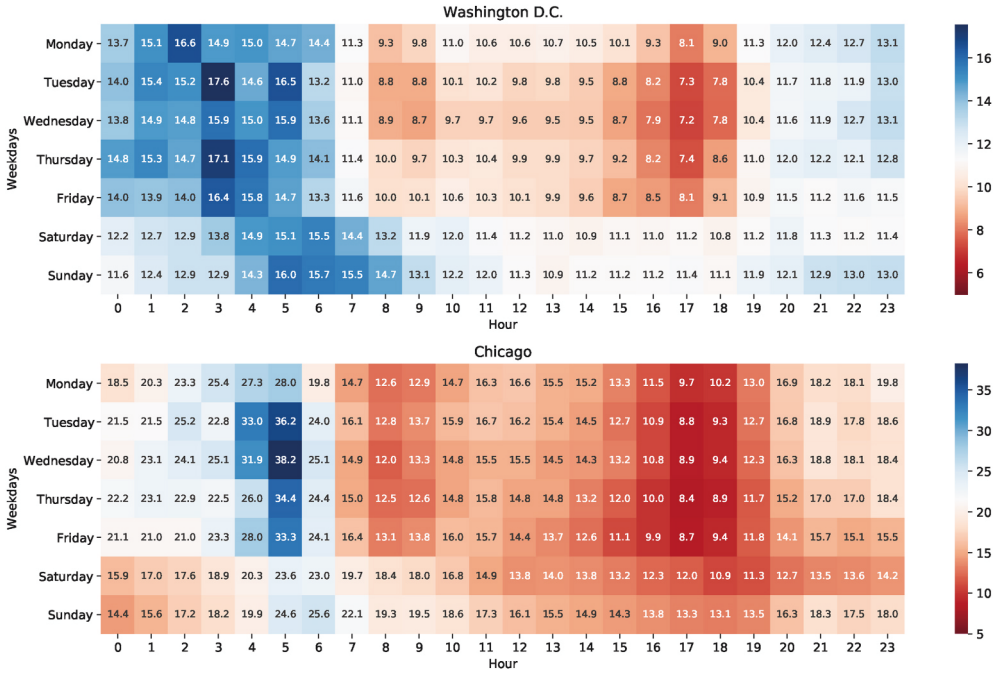

Fig. 1. Heatmaps for the taxi travel speeds in D.C. and Chicago.

eliminating the need for conversions. We collected similar rail transit trip information from GM as we had done for rail transit travels in D.C. In addition, a label was added to indicate whether the trip is from Chicago O'Hare International Airport, as the rail fare depends on it. The "L" train fare is flat, $2.5; it increases to $5 if the trip is from O'Hare to downtown.

We also found the discrepancy in the assumed average walking speeds in the WMATA Trip Planner and GM. As the assumed walking speed by WMATA was significantly smaller than the walking speed adopted in GM (estimated through trip query results from GM), we adjusted the access and walking times in D.C. by using the average walking speed of 3 miles per hour.

3.3. Data preparations

For each taxi trip in the dataset after cleaning, the distance between its either trip end (origin or destination) and the nearest rail transit station was calculated. If either distance was over the limit, which is 1 mile in this study, the taxi trip was removed because essentially rail transit was not a viable travel mode given the significant walking needed. The locations of rail transit stations in both cities were obtained from the GTFS (General Transit Feed Specification) data. Fig. 2 shows the names of selected metro lines and stations as well as the D.C. city limits. Fig. 3 shows the distributions of trip ends (origins and destinations) and metro stations before and after the trip filtration by access distance to rail stations. Clearly, the D.C. downtown is served well by the Washington Metrorail.

For the subsequent analyses, we selected taxi trips in two time periods (8 a.m.–10 a.m. and 4 p.m.–7 p.m.) on Monday, Friday and Sunday only. The origin, destination, and trip start time of each taxi trip in selected time periods were used as inputs to the rail transit planners in obtaining rail transit travel information. The maximum walking distance at either end of the travel was 1 mile. All taxi timestamps were in 2017 while the trip planners did not accept past time as departure time. The same time of day, day of week, and month of year were used, while the year 2017 was replaced by 2019. As discussed above, trip planners did not return valid rail transit travel plans for all requests for various reasons. Only those trips with a viable rail transit plan were kept. It is worth mentioning that the trip planner could return a valid travel plan, while

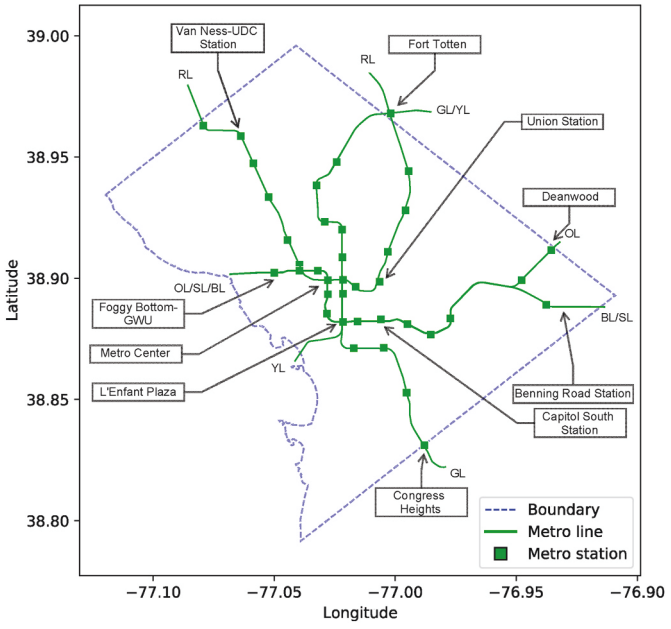

Fig. 2. WMATA Metrorail stations and lines inside of D.C.

rail transit was not involved, i.e., a walking-only trip. Such walking-only trips were removed as well, because the objective of this study is to compare taxi with rail transit, not with walking.

Table 1 shows how the number of trips changes in each stage of the process. In the final dataset, there are 8669 trips in D.C. and 13,635 trips in Chicago. In this study, we assume the maximum walking distance at either trip end is 1 mile. Fig. 4 shows how the number of trips kept in the dataset changes with the walking distance limit at either trip end. When no limit applies, 100% of trips are kept; when the limit is 1 mile, 90% of trips are kept, which means for 90% of the trips, rail transit is a viable transportation mode. When the maximum walking distance is 0.5 miles, around 70% of trips are kept. The main takeaway from Fig. 4 is that in both cities, a considerable proportion (e.g., 70%) of taxi trips can be

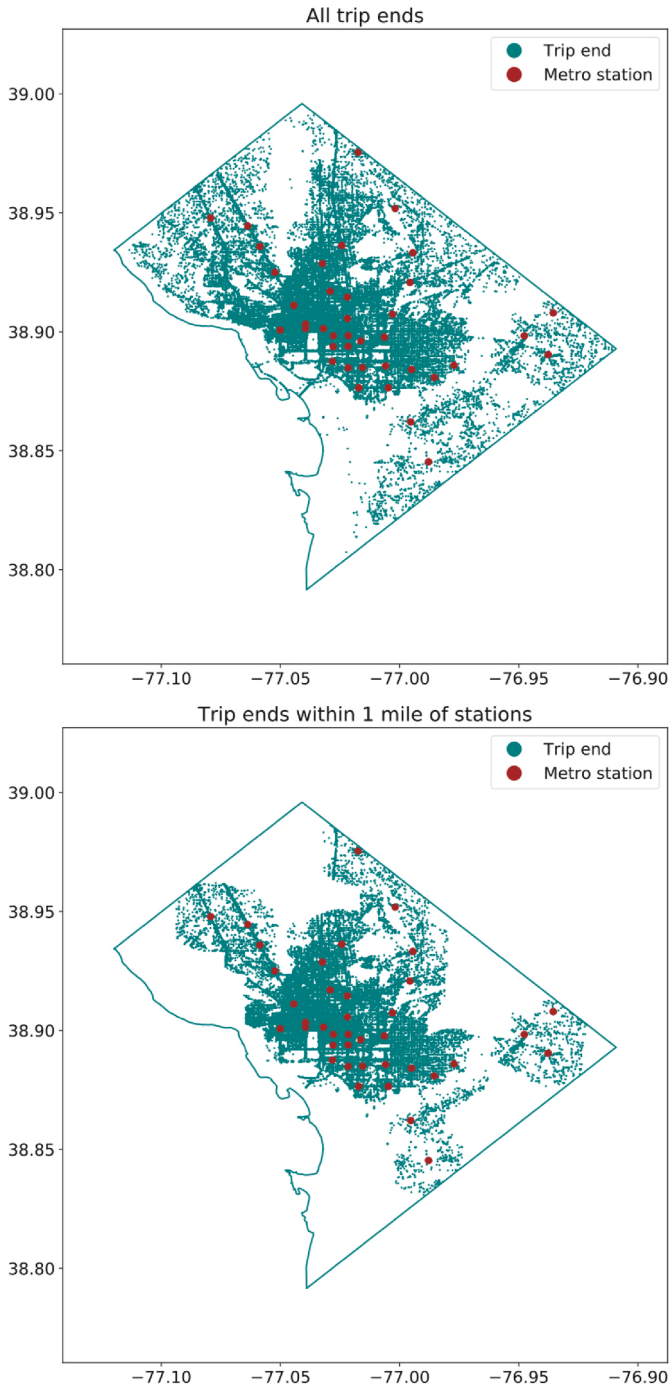

Fig. 3. Distributions of metro stations and trip ends in D.C.

Table 1
Trip number in each stage of analysis.

Stage	D.C.	Chicago
Original	232,510	246,975
After removing trips beyond city limits	178,107	246,975
After data cleaning	92,599	137,283
After filtration by access/egress distance to/from rail stations	82,841	125,584
After selection by time of day and day of week	8966	15,318
After filtration by responses from rail trip planners	8669	13,635

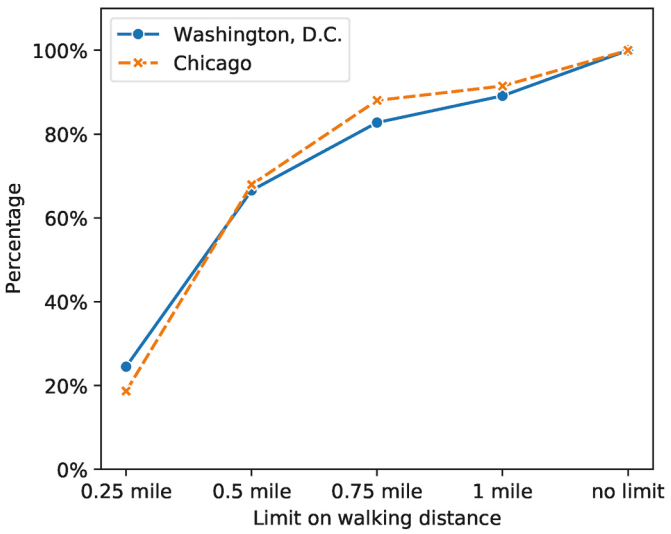

Fig. 4. Effect of rail transit access/egress distance.

substituted by rail transit under reasonable assumptions about the maximum walking distance at either trip end (e.g., 0.5 miles).

4. Methodology

4.1. Metric design

For ease of presentation, we use i to denote an observed trip or simply an observation. For observation i, t_i^{taxi} is the travel time by taxi; c_i^{taxi} is the travel cost by taxi; t_i^{rail} is the travel time by rail transit; c_i^{rail} is the travel cost by rail transit. For each trip i, we then define travel gradient θ_i in Eq. (1) to quantify the comparative advantage of one mode over the other:

$$\theta_i = \frac{c_i^{taxi} - c_i^{rail}}{t_i^{taxi} - t_i^{rail}} \tag{1}$$

The numerator $c_i^{taxi} - c_i^{rail}$ is the cost difference Δ_c between traveling by taxi and by rail transit, which is also the relative cost of taxi to rail transit (assuming the cost of travel by rail transit is 0). The denominator $t_i^{taxi} - t_i^{rail}$ is the travel time difference Δ_t between traveling by taxi and by rail transit, which is also the relative travel time of taxi to rail transit (assuming the travel time by rail transit is 0). Depending on the signs of Δ_c and Δ_t, we classify θ_i into four categories: (I) $\Delta_c > 0$ and $\Delta_t > 0$; (II) $\Delta_c > 0$ and $\Delta_t < 0$; (III) $\Delta_c < 0$ and $\Delta_t < 0$; and (IV) $\Delta_c < 0$ and $\Delta_t > 0$.

As shown in Fig. 5, for each trip we first plot one node (shown as a

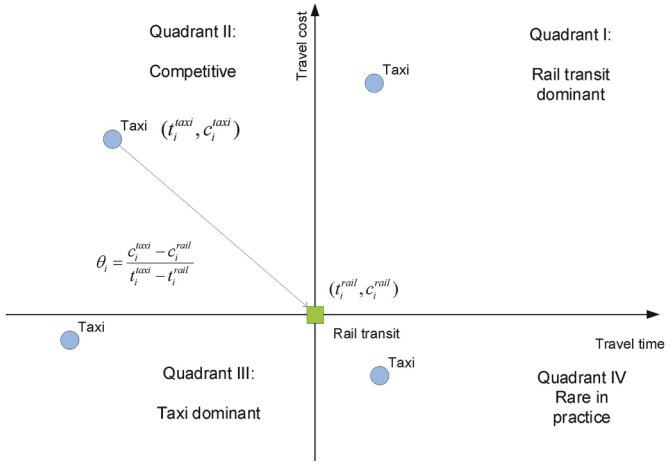

Fig. 5. Quantification of rail transit advantage over taxi for one trip.

square) for the rail transit travel as the origin with coordinates (t_i^{rail}, c_i^{taxi}) on the travel time – cost coordinate plane. Another node for the taxi travel (shown as a circle) with coordinates (t_i^{rail}, c_i^{rail}) can be plotted on the same coordinate plane. The travel time and travel cost axes divide the plane into four quadrants. The node for the taxi travel could fall in one of the four quadrants. If the taxi node falls in Quadrant I, i.e., $\Delta_c > 0$ and $\Delta_t > 0$, the taxi travel has a larger cost and a longer travel time, which means rail transit is the dominant mode. If the taxi node falls in Quadrant III, i.e., $\Delta_c < 0$ and $\Delta_t < 0$, the taxi travel has a smaller cost and a shorter travel time, which means taxi is the dominant mode. If the taxi node falls into Quadrant II, taxi is more expensive, while it saves travel time, which means rail transit is competitive with taxi. If the taxi node falls in Quadrant IV, taxi is less expensive while more time-consuming, which is very rare in practice.

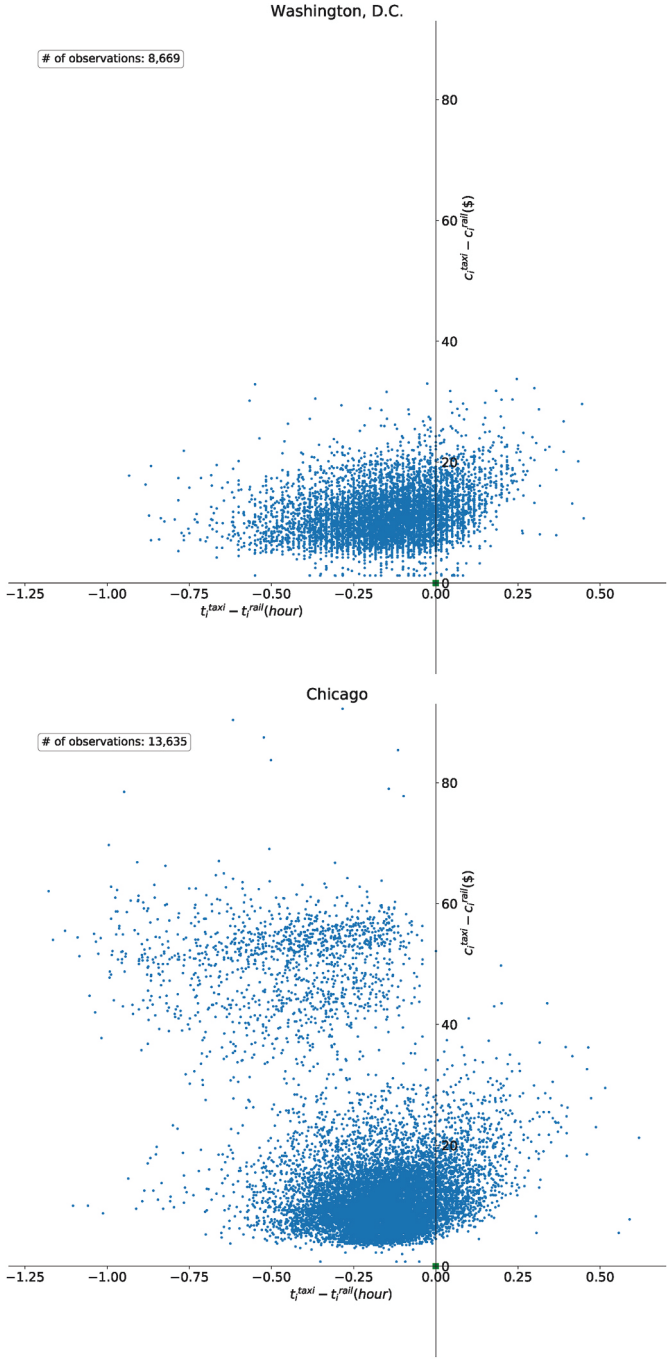

Fig. 6. Visualizations of relative travel costs and travel times of taxi.

Since such a plot can be generated for each trip, such plots can be overlaid to obtain scatter plots of taxi nodes in one figure, as shown in Fig. 6. The distribution of taxi nodes in such a figure visualizes the comparative advantage by considering all observed trips.

Fig. 6 shows that for both cities, there are no taxi nodes in Quadrants III and IV, because for any trip, taxi is more expensive than rail transit. Since there are no nodes in Quadrant III, we conclude taxi dominates rail transit for none of the trips. A significant number of taxi nodes fall into Quadrant I, which means if the taxi trip is undertaken by rail transit, both the travel cost and travel time drop. In particular, 13.7% of taxi nodes belong to Quadrant I in D.C., and 10.5% of taxi nodes belong to Quadrant I in Chicago. Two clusters can be observed for the Chicago taxi trips, because airport-related trips were included only in the Chicago dataset. Note that in D.C. airport-related trips are not included because none of three major airports in the D.C. area are located within the D.C. limits.

4.2. Economic interpretation

As expected, Quadrant II has the most taxi nodes, which means for the majority of trips, taxi and rail transit are competitive: rail transit is less expensive and taxi is less time-consuming. For every trip i in Quadrant II, the value of its gradient θ_i is can be interpreted as the marginal travel cost change due to mode switching from taxi to rail transit. More intuitively, θ_i measures how much money can be saved if this trip is undertaken by rail transit rather than taxi, if the traveler is willing to accept a prolonged travel time by one additional unit of time (e.g., hour).

The concept of gradient θ_i can be better illustrated through rearranging the its original definition in Eq. (1) as follows:

$$\overline{\theta}_i = \frac{\overline{c}_i^{rail} - c_i^{taxi}}{\overline{t}_i^{rail} - t_i^{taxi}} \tag{2}$$

For a trip i that is undertaken by taxi, the travel time is t_i^{taxi} and the travel cost is c_i^{taxi}. Suppose that the trip is now undertaken by rail transit whose travel cost is $\overline{c}_i^{rail}$ while its travel time is exactly 1 h longer than t_i^{taxi}, i.e., $\overline{t}_i^{rail} = t_i^{taxi} + 1$. Clearly, the denominator of Eq. (2) becomes 1; we then obtain $\overline{\theta}_i = \overline{c}_i^{rail} - c_i^{taxi}$, which is the cost saving due to mode switching. For trip i in Quadrant II, its gradient $\overline{\theta}_i$ is negative, which means a positive cost saving is expected.

Considering the tradeoff between cost and time, if the absolute value of $\overline{\theta}_i$ is larger than one traveler's value of time, the mode switch is in general worthwhile. The larger the absolute value of $\overline{\theta}_i$ is, the more worthwhile the mode switch is.

4.3. Regression analyses

After quantifying the comparative advantage with travel gradient, we explore the relation between various trip characteristics (e.g., transit access distance, taxi mileage, and day of week) and travel gradient θ, with and without considering the spatial variations of such a relation. In a global regression model, it is assumed that the relation is independent of location and all the observed data at different locations are used to estimate the same set of parameters. The disadvantage of a global regression model is that the effect of regional factors cannot be captured. For instance, the effect of walking for 0.5 miles in urban cores should be very different from the same amount of walking in suburban areas. A global regression model may potentially obscure the underlying relation between location and travel gradient. In contrast, a local regression model assumes location-specific relations. Only the data observed at or around a specific location are used to estimate such local relations. Therefore, the estimated relations could vary over space. In this study, we use Multiple Linear Regression (MLR) and Geographically Weighted Regression (GWR) to conduct global and local regressions, respectively.

4.3.1. MLR model

The MLR model takes the following form:

$$\theta_i = \beta_0 + \sum_k \beta_k x_{ik} + \varepsilon_i, \forall i \quad (3)$$

where θ_i is the ith observation of travel gradient, x_{ik} is the ith observation of the kth independent variable, and ε_i is a normally distributed error term with zero mean. It is typically assumed in a linear regression analysis that there are no correlations among independent variables or between two successive observations of the same variable, i.e., no multicollinearity or autocorrelation. Each β_k as well as β_0 need to be estimated. In MLR, we consider transit access distance (distance from the trip origin to the nearest transit station), transit egress distance (distance from the last transit station to the trip destination), taxi mileage, number of transit transfers, time period of day, day of week, and airport trip indicator (0 or 1) as independent variables, primarily based on the data availability. Since both time of day and day of week are categorical variables, to avoid a linear dependency, we leave out one of these indicators. For example, we consider Monday, Friday, and Sunday, while including only two indicator variables for Friday and Sunday, respectively.

4.3.2. GWR model

The Geographically Weighted Regression (GWR) model can consider the spatial information of observed data and explore the spatially varying relationships between dependent and independent variables. Eq. (4) represents a GWR model where coefficients of independent variables are specific to geographical coordinate (u_i, v_i). In contrast, coefficients in MLR are location independent.

$$\theta_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i, \forall i \quad (4)$$

GWR builds a regression model for each location where sample i is observed. To estimate the coefficients, a buffer is used to determine what neighboring observations are considered in the coefficient estimation. This buffer is also known as a spatial kernel. Generally, all observations in the buffer are weighted differently according to their distance to the location of observation i. The further away an observation is, a smaller weight it carries. When an observation falls out of the kernel (its distance to observation i exceeds the bandwidth), its weight is zero.

There are many choices of kernels. A fixed kernel includes all observations that are within a certain distance (a fixed bandwidth) from the point of interest. When the spatial distribution of data is relatively uniform, a fixed kernel can be chosen; otherwise, an adaptive kernel is more commonly used where the number of neighboring observations is optimizable (Cheng et al., 2021; Zhao and Cao, 2020; Baker, 2020). In this study, an adaptive kernel is used. The shape of a kernel, which determines how different observations are weighted, could be uniform, Gaussian, exponential, bisquare, and tricube, among others. We use a bisquare kernel, which is commonly used in the literature (Chiou et al., 2015; Munira and Sener, 2020; Cordera et al., 2019).

It is also understandable that the choice of kernel bandwidth has a major influence on GWR results, because when the bandwidth is too small, not enough observations are used to estimate the coefficients, implying large errors; when the bandwidth is too large, a local regression reduces to a global one, thus masking local variations. We follow the bandwidth optimization framework proposed in Brunsdon et al. (1996) and Brunsdon et al. (1998), which is described as follows. Let h be a bandwidth and $\widehat{\theta}_i(h)$ be the predicted value of θ_i by GWR for the given bandwidth h. We further introduce $\widehat{\theta}_{\neq i}(h)$ as the GWR-predicted value of θ_i with the observations for location i omitted. Then, the optimal bandwidth is found by minimizing the cross-validated (CV) sum of squared errors (Zhong and Li, 2016), defined as follows:

$$CV = \sum_i [\theta_i - \widehat{\theta}_{\neq i}(h)]^2 \quad (5)$$

GWR has been implemented in many programming languages. In this study, a R package named GWmodel (Gollini et al., 2015) is used.

5. Results

5.1. Descriptive statistics

In this section we focus on Quadrant II trips only, which account for 86.3% of trips in D.C. and 89.5% of trips in Chicago. For such trips in D.C., both the average transit access distance and transit egress distance are 0.37 mile; in Chicago, the average access distance is 0.3 mile, and the egress distance is 0.35 mile. Since the absolute value of θ can be extremely large when the travel time difference Δ_t, which is the denominator, is close to zero, we further filtered trips by travel time difference and kept only such trips with a rail transit travel time at least 20% larger than the taxi travel time, i.e., $(t_i^{rail} - t_i^{taxi})/t_i^{taxi} > 20\%$. Fig. 7 shows the distributions of θ for all trips contained in the final dataset. The average value of θ is -$76.17/hour in D.C., which means if a traveler can accept a trip prolonged by 1 h, a trip cost reduction of $76.17 is expected. This hourly travel cost saving outnumbers the value of time for most travelers. The average value of θ is -$68.79 in Chicago, which can be interpreted in the same manner. As θ is used as the response variable in regression analyses, we take the square root of $-\theta$, so that the regression models fit the data better.

After checking the correlation matrix, we find explanatory variables are generally weakly correlated with a correlation coefficient smaller than 0.15. The highest correlation efficient is around 0.4, which is between the number of transfers and taxi mileage. When higher correlations exist, one of the correlated variables should be removed.

5.2. MLR results

Tables 2 and 3 show the linear regression results for D.C. and Chicago, respectively. For continuous predictor variables (namely access distance, egress distance, total mileage and the number of transfers), a coefficient represents the change in the predicted value of the response variable for each one-unit change in a predictor variable, if other predictor variables remain constant. The negative coefficients of access distance, egress distance, and the number of transfers indicate that when values of such predictors increase, the marginal travel cost saving due to mode switch from taxi to rail transit decreases, which is understandable. When the access/egress distance and the number of transfers grow, rail transit becomes more time-consuming and less competitive. For both cities, the total mileage by taxi has a positive coefficient, implying an increasing marginal cost saving due to an increased total mileage. This positive effect is significant in Chicago, while it is not as significant in D.C.

A categorical variable has multiple levels, with only one level selected as the reference. For other indicator variables, the coefficient measures how the response value changes when the level of the categorical variable changes from the reference level to the level associated with the indicator variable. Regarding the day-of-week variable, with Monday as the reference point, the marginal travel cost saving is higher on Friday and lower on Sunday. A comparison of p-values indicates that the indicator variable for Friday is not as significant as the indicator variable for Sunday in both cities, meaning that on Sunday the comparative advantage (marginal cost saving) diminishes substantially. For the afternoon trips, the marginal travel cost saving is higher than trips in the morning, with 8 a.m. as the reference. The relatively large p-values for the indicator variable for 9 a.m. in both cities indicate that the change in the marginal travel cost saving is insignificant when the time of travel changes from 8 a.m. (reference time of day) to 9 a.m. Results also show that in Chicago travelers can save significantly if they switch from taxi to rail transit when their travels are airport-related.

The intercept shown in Table 2 is the value of transformed dependent

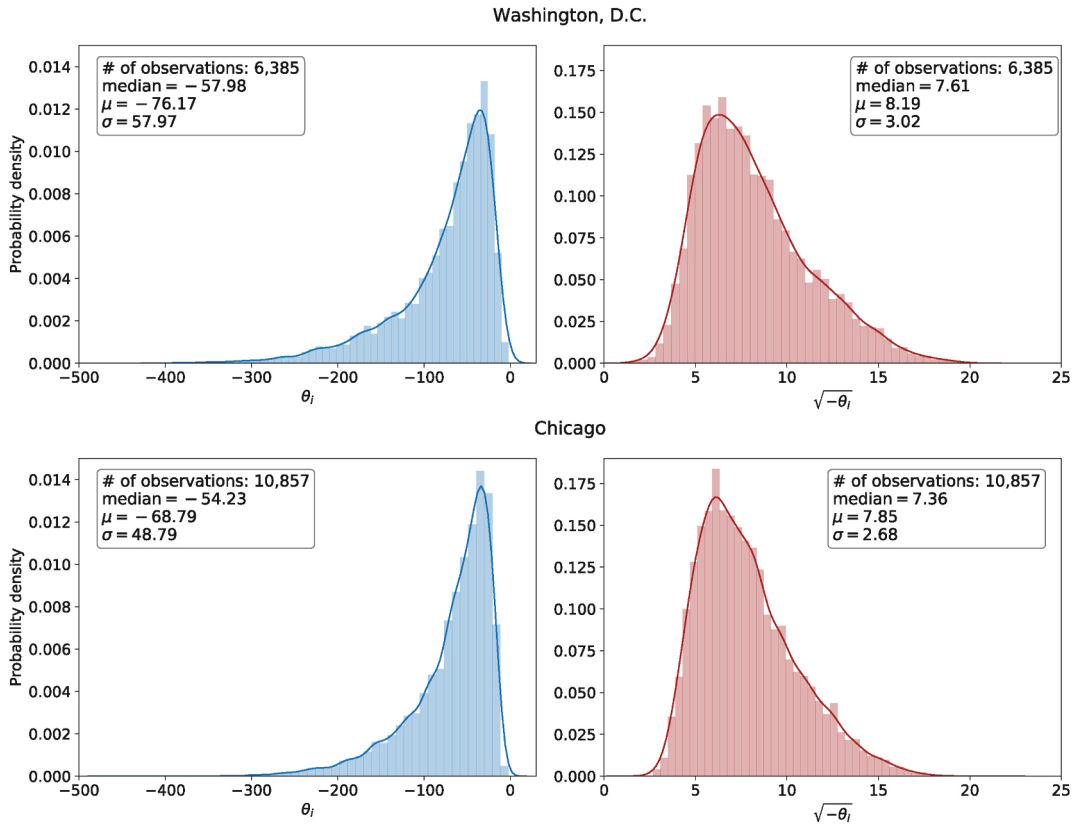

Fig. 7. Distributions of trip gradients and their transformations.

Table 2
MLR results for Washington, D.C.

Independent variable	Washington, D.C.				
	Estimate	Std. Error	t value	*p*-value	95% confidence interval
(Intercept)	12.148	0.128	94.826	0.000***	(11.897, 12.400)
Access distance	−5.125	0.150	−34.204	0.000***	(-5.419, −4.832)
Egress distance	−4.697	0.143	−32.905	0.000***	(-4.978, −4.418)
Total distance	0.079	0.024	3.257	0.001**	(0.032, 0.127)
No. of transfer	−2.163	0.076	−28.477	0.000***	(-2.312, −2.014)
Hour					
8 a.m. (reference)					
9 a.m.	0.150	0.116	1.293	0.209	(-0.078, 0.379)
4 p.m.	0.499	0.110	4.525	0.000***	(0.283, 0.716)
5 p.m.	0.715	0.110	6.486	0.000***	(0.499, 0.932)
6 p.m.	0.479	0.109	4.385	0.000***	(0.265, 0.694)
Day of week					
Monday (reference)					
Friday	0.130	0.073	1.790	0.073*	(-0.012, 0.273)
Sunday	−0.700	0.083	−8.391	0.000***	(-0.864, −0.537)

Note: * = $p \leq 0.10$; ** = $p \leq 0.05$; *** = $p \leq 0.001$.

Table 3
MLR results for Chicago.

Independent variable	Chicago				
	Estimate	Std. Error	t value	*p*-value	95% confidence interval
(Intercept)	9.346	0.094	99.533	0.000***	(9.162, 9.530)
Access distance	−3.789	0.131	−28.831	0.000***	(-4.047, −3.531)
Egress distance	−3.539	0.130	−27.231	0.000***	(-3.794, −3.284)
Total distance	0.315	0.007	42.294	0.000***	(0.301, 0.330)
No. of transfer	−1.65	0.056	−29.522	0.000***	(-1.760, −1.541)
Hour					
8 a.m. (reference)					
9 a.m.	0.067	0.074	0.905	0.366	(-0.078, 0.211)
4 p.m.	0.462	0.073	6.371	0.000***	(0.320, 0.604)
5 p.m.	0.630	0.074	8.515	0.000***	(0.485, 0.775)
6 p.m.	0.397	0.071	5.563	0.000***	(0.257, 0.537)
Day of week					
Monday (reference)					
Friday	0.145	0.053	2.741	0.006**	(0.041, 0.249)
Sunday	−0.418	0.056	−7.500	0.000***	(-0.527, −0.309)
Airport trip					
No (reference)					
Yes	1.294	0.145	8.897	0.000***	(1.009, 1.578)

Note: * = $p \leq 0.10$; ** = $p \leq 0.05$; *** = $p \leq 0.001$.

variable when all continuous variables are set to be zero and all categorical variables take the reference value. As setting all independent variables to be zero may not represent any realistic scenario, interpreting intercept is not as useful as interpreting other coefficients in analyzing the relations between independent variables and the transformed dependent variable. In addition, the last column of Table 2 shows the 95% confidence interval for each coefficient to be estimated.

For both linear regression models, the multiple R^2 and adjusted R^2 values are close to 0.3, which is relatively low. However, as *p*-values

indicate that most selected explanatory variables are significant. That means if the purpose of the MLR model is to make predictions, the accuracy of the model is low, as the assumed linear relation cannot explain much of the variation. Nonetheless, low *p*-values indicate the observed relation between selected explanatory variables and the response variable is statistically significant, which satisfies the need in this study.

5.3. GWR results

We further run GWR on the D.C. data only, because trip origins and destinations were approximated by centroids of census tracks and community areas in Chicago, leaving the number of unique geographic locations (origins and destinations) to be only 148. The relatively small number of locations and highly uneven distribution of locations in the Chicago dataset mean that when a regression model is fitted for some location, it is likely that too few nearby locations could be covered by the spatial kernel. Therefore, the few observations used for fitting a regression model could not yield robust coefficient estimates. The resulting goodness of fit is thus quite low in the case of Chicago.

As each observation i is associated with two geographic locations (origin and destination), we develop two GWR models, one based on origin and the other on destination. The optimal numbers of nearest neighbors for those two GWR models are 672 and 701, respectively. For each GWR model, a regression is built for each location, which has its location-specific coefficient estimations. Therefore, we can obtain the distributions of coefficients of each independent variable, which will reveal how the impact of an independent variable varies across different locations. Table 4 provides the basic statistics (minimum, maximum and median) of the coefficients of each independent variable for both the origin-based and destination-based GWR models.

For each location-specific regression model, the *p*-value for an independent variable can be compared with the significance level (0.05) to check whether the independent variable is statistically significant. Table 5 shows the percentage of significant coefficients of each independent variable for both GWR models. Among all statistically significant coefficients, we further show the sign of a coefficient (positive or negative) in Table 5. We can find that for most location-specific regression models, access distance, egress distance, number of transfers, and the indicator variable for Sunday are in general statistically significant. For such variables, we can also observe that the sign of the coefficient is definite (100% positive or negative, rather than a mix). This further implies the effect (positive or negative) of such variables on travel gradient is constant over space, although the magnitude varies. In contrast, total distance by taxi has a low percentage of statistically significant coefficients. The sign of its coefficient is indefinite, as indicated further in Fig. 8. Clearly, we can see from Fig. 8 that the coefficient of taxi mileage ranges from −0.5 to 1.2 in the origin-based model and from −0.6 to 0.6 in the destination-based model. This means the effect of taxi mileage on travel gradient could be positive or negative, depending on locations.

Table 4
GWR results for Washington, D.C.

Independent variable	Origin-based GWR			Destination-based GWR		
	Min.	Median	Max.	Min	Median	Max.
(Intercept)	8.465	12.427	15.147	9.009	12.534	14.809
Access distance	−13.194	−5.887	−0.805	−8.680	−5.311	−2.908
Egress distance	−8.102	−5.163	−1.461	−8.738	−5.370	−1.621
Total distance	−0.559	0.081	1.187	−0.636	0.0722	0.637
No. of transfer	−3.813	−2.267	−0.589	−3.503	−2.257	−0.592
Hour						
8 a.m. (reference)						
9 a.m.	−2.228	0.132	1.702	−1.282	0.153	1.255
4 p.m.	−0.806	0.346	0.838	−0.672	0.521	1.977
5 p.m.	−0.486	0.794	3.339	−0.314	0.794	2.038
6 p.m.	−1.352	0.589	2.760	−0.718	0.442	1.742
Day of week						
Monday (reference)						
Friday	−0.930	0.130	1.055	−0.572	0.135	0.875
Sunday	−1.890	−0.805	0.211	−1.627	−0.764	0.456

We then analyze the spatial variations of the coefficients of selected independent variables. Fig. 9 shows how the coefficient of transit access distance varies over space. Although all coefficients are negative, the coefficient is less negative in the Union Station area in the origin-based GWR model; the coefficient is less negative in two notable areas (near Fort Totten in the north and Congress Heights in the south) in the destination-based GWR model. A less negative coefficient means the response variable decreases less quickly as the subject independent variable increases. In other words, the drop in the marginal cost saving is not very sensitive to the increase in transit access time for people traveling from the Union Station area as well as people traveling to the Fort Totten and Congress Heights areas.

Fig. 10 shows how the coefficients of the indicator variable for 5 p.m. vary over space. As shown in Table 4, the coefficients for 5 p.m. are positive at most locations, which means as compared to 8 a.m., traveling at 5 p.m. by rail transit yields more cost savings. Specifically, the savings are comparatively more evident for travels from the National Mall (especially near Capital South Station shown in Fig. 10(a)) and to the northwestern region of D.C. (especially near Van Ness-UDC Station in Figure Fig. 10(b)).

Similarly, from Fig. 11 and Table 4, we conclude that rail transit travels on Sunday yield comparatively smaller cost savings with Monday as the reference point. Nonetheless, the decrease in cost savings is not evident for travels from the Congress Heights area (as shown in Fig. 11 (a)) and other travels to the Capital South area (as shown in Fig. 11(b)).

The average R^2 values of the origin and destination-based GWR models are 0.40 and 0.38, respectively. The R^2 values are higher than the R^2 value of the MLR model, meaning that after further considering locations, the regression model can explain the variation of the dependent variable (travel gradient) better. Fig. 12 shows how the R^2 value varies over space.

In summary, as GWR can capture the variations of the relation between various trip characteristics and travel gradient over space, a higher goodness of fit is achieved than MLR. More importantly, GWR reveals highly promising travel directions and time periods for targeted rail transit marketing.

6. Discussions

6.1. Policy implications

This comparative study of two urban transportation modes based on the empirical data collected in Washington, D.C. and Chicago yields a few important findings. First, we find approximately 70% of taxi trips can be substituted by rail transit if the maximum walking distance is 0.5 miles. This is largely consistent with Wang and Ross (2019), who classified approximately 60% of taxi trips in New York City as transit-competing using a different criterion. Second, if a taxi rider can accept an increased travel time by 1 h, she/he can reduce the travel cost by around $70 on average, exceeding the value of travel time for most travelers. Clearly, if the trip is prolonged by half an hour due to the mode switch, at least half of $70 (namely $35) can be saved. The travel cost reduction is slightly higher in Washington, D.C. than in Chicago. Third, in both cities rail transit dominates taxi for around 10% of the taxi trips, while taxi never dominates rail. For the rest of trips, rail transit is competitive with rail and significant travel cost reductions are achievable if prolonged travel times are acceptable. The last two findings have not yet been reported in the literature.

While taxi and rail transit are supposed to serve different segments of urban travelers given their mode-specific characteristics, it is understandable that both modes are considered as attractive options to a

Table 5
Statistical significance of the relationships derived from GWR.

Independent variable	Origin-based GWR			Destination-based GWR		
	% of significance (p<0.05)	Among significant		% of significance (p<0.05)	Among Significant	
		Positive	Negative		Positive	Negative
Intercept	100.0%	100.0%	0.0%	100.0%	100.0%	0.0%
Access distance	99.7%	0.00%	100.0%	100.0%	0.0%	100.0%
Egress distance	99.9%	0.0%	100.0%	100.0%	0.0%	100.0%
Total distance	38.9%	62.2%	37.8%	29.9%	72.2%	27.7%
No. of transfer	98.5%	0.0%	100.0%	100.0%	0.00%	100.00%
Hour						
8 a.m. (reference)						
9 a.m.	2.7%	67.6%	32.3%	4.9%	93.6%	6.3%
4 p.m.	20.2%	100.0%	0.0%	23.0%	100.0%	0.0%
5 p.m.	41.4%	100.0%	0.0%	45.7%	100.0%	0.0%
6 p.m.	22.4%	91.5%	8.4%	25.4%	100.0%	0.0%
Day of week						
Monday (reference)						
Friday	18.0%	90.77%	9.2%	12.1%	99.3%	0.6%
Sunday	59.0%	0.0%	100.0%	70.4%	0.0%	100.0%

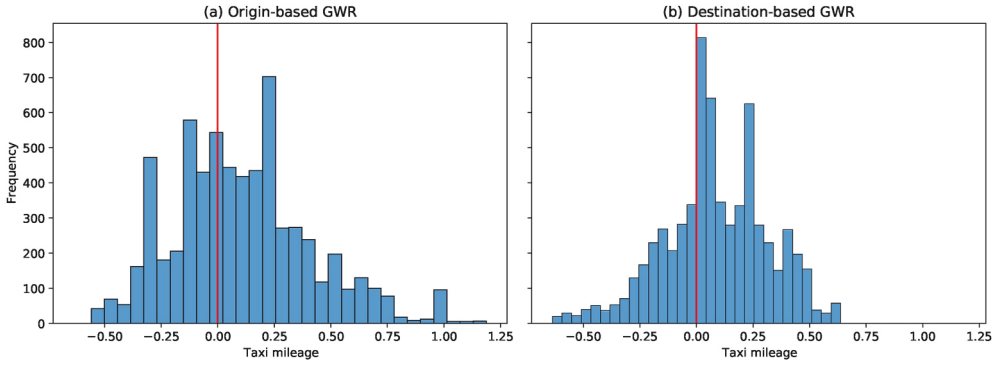

Fig. 8. Histograms of coefficients of taxi mileage in two GWR models.

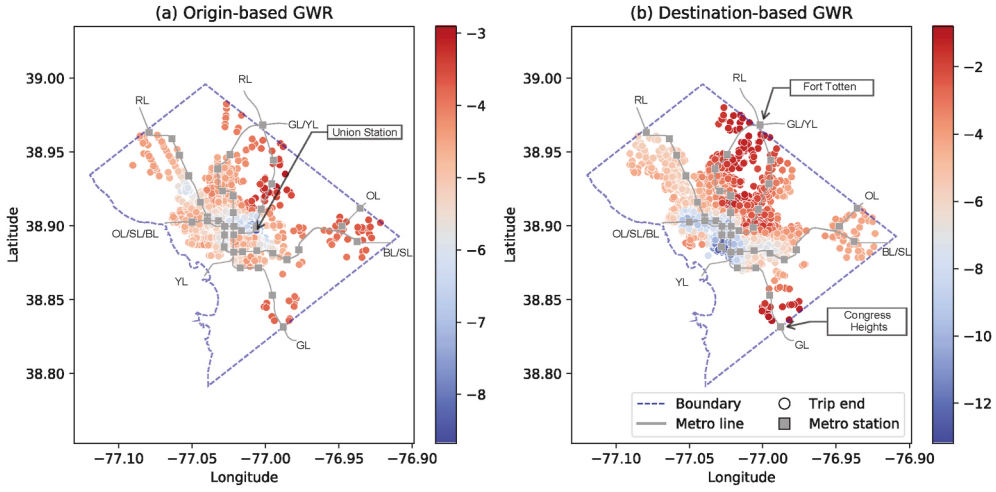

Fig. 9. Variations of the coefficient of transit access distance.

certain group of travelers, representing the overlap between those two segments of travelers. Due to the high competitiveness of rail transit identified through this study, especially in certain regions and time periods, the main takeaway for rail transit operators is that proper rail transit marketing campaigns can be launched to increase rail transit ridership by highlighting the relative advantage of rail transit over taxi. Since the two GWR models have generated important information about the spatial variations of the relation between a trip characteristic and travel gradient, rail transit marketing could be targeted at specific locations and time periods. For instance, for travels to the Capitol Heights area, the marginal travel cost saving is substantial despite increases in transit access time, if taxi substituted by rail transit. Therefore, such corresponding travelers could be targeted by the marketing campaign. To ensure the efficacy of the rail transit marketing campaign, the way to delivery such marketing messages to customers should be carefully evaluated and selected (Hess and Bitterman, 2016; Andersson et al.,

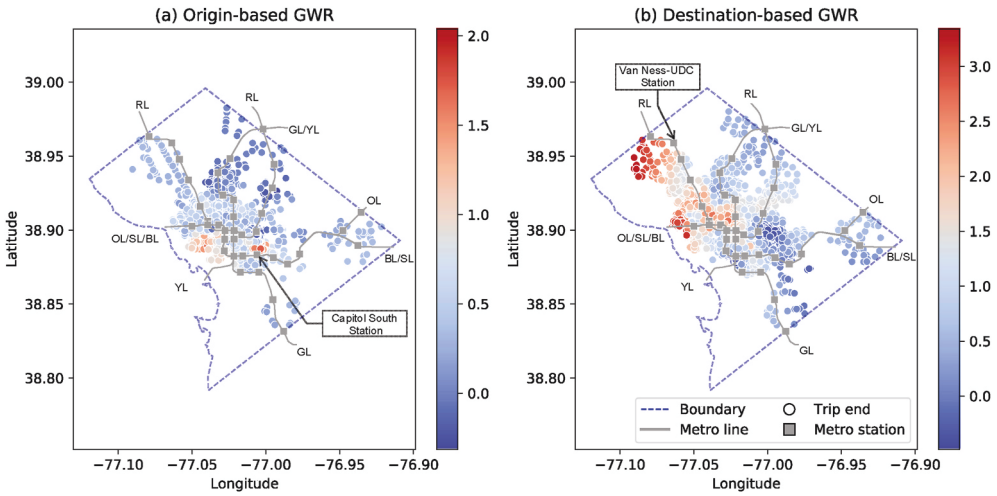

Fig. 10. Variations of the coefficient of the indicator variable for 5 p.m.

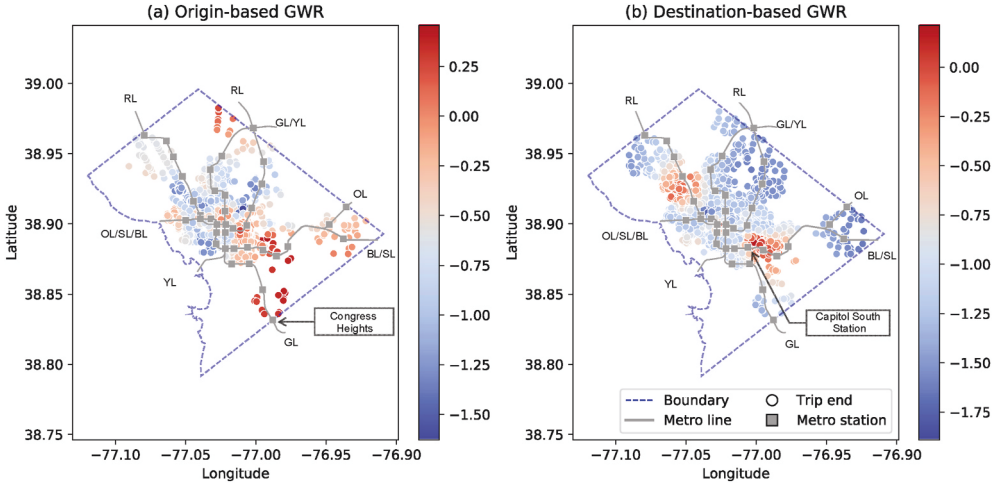

Fig. 11. Variations of the coefficient of the indicator variable for Sunday.

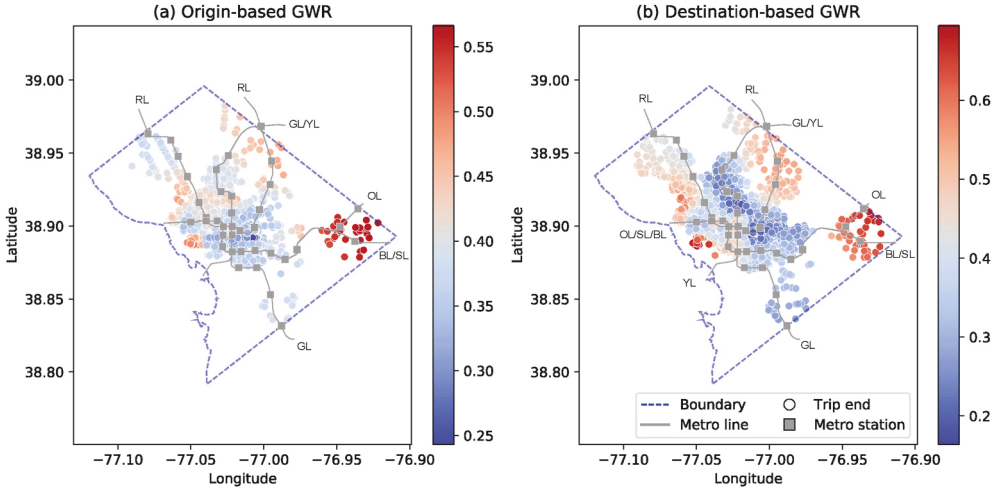

Fig. 12. Local R^2 values.

2020). For instance, the following dynamic message can be used: "Do you know that if you take metro instead of taxi to Capitol Heights, you can save $35 while you only spend extra 27 min relaxing on a metro train?" Such a message can be customized to trip origins, destinations, as well as time periods, based on the GWR results. As such messages can also be personalized, they can be pushed through mobile APPs and other

web portals, instead of being printed on posters. The improved marketing plan by highlighting the unique advantage of rail transit could thus potentially increase rail transit ridership and promote transportation sustainability.

For the overall efficiency and sustainability of urban transportation systems, taxi should avoid directly competing with rail transit for customers, as the latter is widely believed to consume less energy and reduce air pollution (Ghimire and Lancelin, 2019). Therefore, transportation regulatory authorities should enact those policies that give priority to the development of more sustainable and efficient transportation modes and mitigate their direct competition with other modes.

6.2. Limitations of this study

In this study, rail transit and taxi are compared by travel cost and time only, while other trip attributes, such as comfort level, reliability and environmental impact, are not considered yet. For certain riders, depending on their demographic characteristics, taxi may be the only choice (Wong et al., 2020). For example, individuals with disabilities cannot take rail transit and the tradeoff between travel cost and travel cost does not play a major factor in their mode choices. Another example is that some passengers who carry heavy luggage or have a very tight schedule may also consider taxi only. Therefore, to enable a more comprehensive mode comparison, other service attributes of transportation modes and the demographic characteristics of urban travelers should be further incorporated. It is worth noting that Ulak et al. (2020) estimated the value of convenience (which was close to $30) based on taxi trip data in NYC. It was argued by Ulak et al. (2020) that the main competitiveness of taxi in NYC is attributed to convenience.

The taxi-only origin of the trip data presents potential statistical bias. To compare two modes fairly, namely rail transit and taxi, trip data (including trip origin, destination, departure time, etc.) in the urban environment should be randomly generated and used for comparison. As rail transit cannot cover all the urban areas, it is thus possible that taxi becomes the dominant mode for certain trips, due to the unavailability of rail transit. Those trips thus fall in Quadrant III. In contrast, all trips in this study are actual taxi trips, which means taxi has some implicit advantage, because for such trips, travelers actually chose taxi over other modes. Similarly, if all such trips are actual rail transit trips, rail transit would have been given some implicit advantage. It should be noted that detailed rail transit trip data are less widely accessible than taxi trips data.

It is also known that passengers perceive in-vehicle travel time and out-of-vehicle travel time (such as waiting, walking and transferring) differently. Nonetheless, in this study, the total travel time obtained from rail transit trip planners was directly compared with the taxi travel time, without distinguishing in-vehicle travel time from out-of-vehicle travel time. This simplification may potentially underestimate the burden of rail transit travel, as out-of-vehicle travel time is generally more burdensome. In addition, the effect of potential in-vehicle crowding on passengers' perceptions of rail transit travel time was not included in the present study.

7. Conclusions

This study has assessed the comparative advantage of rail transit over taxi using publicly available data in two U.S. cities. We first collect and clean taxi trip data and rail transit travel information. A metric named travel gradient is proposed for quantifying the comparative advantage and its economic interpretation is provided. We also explore how various trip characteristics are related to travel gradient initially using multiple linear regression. To understand the variations of the relations between the trip characteristics and travel gradient at different locations, we then analyze the data with geographically weighted regression. We have derived a few important research findings and analyzed the policy implications in Section 6. Although empirical data are from two different cities, we find very similar results, especially in the percentage of taxi trips that can be substituted by rail transit and the marginal travel cost savings due to mode switching to rail transit. Therefore, the derived research findings should be applicable to other cities with comparable urban rail transit systems. Similar analyses can be conducted for those cities before relevant rail transit marketing policies are designed and enacted.

The current study can be improved in the following ways:

1. As the group size of passengers in a taxi trip is unknown, it is assumed to be 1 for convenience. When multiple passengers share the taxi (Sun and Zhang, 2018), the relative advantage of rail transit over taxi may change considerably. For rail transit, the access/egress time is considered, while for taxi, the waiting time, which can be considered as "taxi access time", is not included, because no taxi waiting data are available. Thus, it is fair to add the waiting time as part of the taxi trip time.
2. All regression models used in this study are for exploring the relations between trip characteristics and travel gradient, rather than for making predictions accurately. If the study purpose is to predict travel gradient, other possible explanatory variables, especially socio-demographic variables that are commonly available in household travel surveys (Ha et al., 2020) as well as built environment variables Hochmair (2016), should be added to the model to improve the model's goodness of fit.
3. Due to the highly uneven distribution of locations in the Chicago dataset, we did not run the GWR for Chicago. More advanced local regression models that can address this issue could be further adopted.

CRediT author statement

Sajeeb Kirtonia: Investigation, Formal Analysis, Writing - Original Draft, Visualization.

Yanshuo Sun: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing.

Acknowledgements

Helpful comments from anonymous reviewers are greatly acknowledged. The corresponding author is partially supported by the National Science Foundation (Nos. 2100745 and 2055347).

References

Andersson, A., Hiselius, L.W., Adell, E., 2020. The effect of marketing messages on the motivation to reduce private car use in different segments. Transport Pol. 90, 22–30.

Baker, D.M., 2020. Transportation Network Companies (TNCs) and public transit: examining relationships between TNCs, transit ridership, and neighborhood qualities in San Francisco. Case Stud. Transport Pol. 8, 1233–1246. https://doi.org/10.1016/j.cstp.2020.08.004.

Brunsdon, C., Fotheringham, A.S., Charlton, M.E., 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. Geogr. Anal. 28, 281–298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x.

Brunsdon, C., Fotheringham, S., Charlton, M., 1998. Geographically weighted regression. J. Roy. Stat. Soc. Ser. D (Stat.) 47, 431–443. https://doi.org/10.1111/1467-9884.00145.

Cheng, L., Shi, K., De Vos, J., Cao, M., Witlox, F., 2021. Examining the spatially heterogeneous effects of the built environment on walking among older adults. Transport Pol. 100, 21–30. https://doi.org/10.1016/j.tranpol.2020.10.004.

Chiou, Y.-C., Jou, R.-C., Yang, C.-H., 2015. Factors affecting public transportation usage rate: geographically weighted regression. Transport. Res. Part A Pol. Pract. 78, 161–177. https://doi.org/10.1016/j.tra.2015.05.016.

City of Chicago, 2016. Chicago Taxi Data Released [Online. https://digital.cityofchicago.org/index.php/chicago-taxi-data-released/. (Accessed 1 August 2020).

Cordera, R., Chiarazzo, V., Ottomanelli, M., dell'Olio, L., Ibeas, A., 2019. The impact of undesirable externalities on residential property values: spatial regressive models and an empirical study. Transport Pol. 80, 177–187. https://doi.org/10.1016/j.tranpol.2016.05.007.

Faghih-Imani, A., Anowar, S., Miller, E.J., Eluru, N., 2017. Hail a cab or ride a bike? a travel time comparison of taxi and bicycle-sharing systems in New York City.

Transport. Res. Part A Pol. Pract. 101, 11–21. https://doi.org/10.1016/j.tra.2017.05.006.
Ghimire, R., Lancelin, C., 2019. The relationship between financial incentives provided by employers and commuters' decision to use transit: results from the atlanta regional household travel survey. Transport Pol. 74, 103–113.
Gollini, I., Lu, B., Charlton, M., Brunsdon, C., Harris, P., 2015. GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models. J. Stat. Software 63, 1–50. https://doi.org/10.18637/jss.v063.i17.
Ha, J., Lee, S., Ko, J., 2020. Unraveling the impact of travel time, cost, and transit burdens on commute mode choice for different income and age groups. Transport. Res. Part A Pol. Pract. 141, 147–166. https://doi.org/10.1016/j.tra.2020.07.020.
Hess, D.B., Bitterman, A., 2016. Branding and selling public transit in north America: an analysis of recent messages and methods. Res. Transport. Bus. Manage. 18, 49–56.
Hochmair, H.H., 2016. Spatiotemporal pattern analysis of taxi trips in New York City. Transport. Res. Rec. 2542, 45–56. https://doi.org/10.3141/2542-06.
Irawan, M.Z., Belgiawan, P.F., Tarigan, A.K.M., Wijanarko, F., 2020. To compete or not compete: exploring the relationships between motorcycle-based ride-sourcing, motorcycle taxis, and public transport in the Jakarta metropolitan area. Transportation 47, 2367–2389. https://doi.org/10.1007/s11116-019-10019-5.
Jiang, S., Guan, W., He, Z., Yang, L., 2018. Exploring the intermodal relationship between taxi and subway in Beijing, China. J. Adv. Transport. https://doi.org/10.1007/s11116-019-10019-5, 2018.
Kim, K., 2018. Exploring the difference between ridership patterns of subway and taxi: case study in Seoul. J. Transport Geogr. 66, 213–223. https://doi.org/10.1016/j.jtrangeo.2017.12.003.
Li, L., Wang, S., Li, M., Tan, J., 2018. Comparison of travel mode choice between taxi and subway regarding traveling convenience. Tsinghua Sci. Technol. 23, 135–144.
Liu, Y., Kang, C., Gao, S., Xiao, Y., Tian, Y., 2012. Understanding intra-urban trip patterns from taxi trajectory data. J. Geogr. Syst. 14, 463–483. https://doi.org/10.1007/s10109-012-0166-z.
Liu, X., Gong, L., Gong, Y., Liu, Y., 2015. Revealing travel patterns and city structure with taxi trip data. J. Transport Geogr. 43, 78–90. https://doi.org/10.1016/j.jtrangeo.2015.01.016.
Ma, T., Liu, C., Erdoğan, S., 2015. Bicycle sharing and public transit: does capital bikeshare affect metrorail ridership in Washington, DC? Transport. Res. Record 2534, 1–9. https://doi.org/10.3141/2534-01.
Munira, S., Sener, I.N., 2020. A geographically weighted regression model to examine the spatial variation of the socioeconomic and land-use factors associated with Strava bike activity in Austin, Texas. J. Transport Geogr. 88, 102865. https://doi.org/10.1016/j.jtrangeo.2020.102865.
Pinjari, A.R., Bhat, C., 2006. Nonlinearity of response to level-of-service variables in travel mode choice models. Transport. Res. Record 67–74. https://doi.org/10.1177/0361198106197700109, 1977.
Sun, Y., Schonfeld, P.M., 2016. Schedule-based rail transit path-choice estimation using automatic fare collection data. J. Transport. Eng. 142, 04015037 https://doi.org/10.1061/(ASCE)TE.1943-5436.0000812.
Sun, Y., Xu, R., 2012. Rail transit travel time reliability and estimation of passenger route choice behavior: analysis using automatic fare collection data. Transport. Res. Record 2275, 58–67. https://doi.org/10.3141/2275-07.
Sun, Y., Zhang, L., 2018. Potential of taxi-pooling to reduce vehicle miles traveled in Washington, DC. Transport. Res. Record 2672, 775–784. https://doi.org/10.1177/0361198118801352.
Ulak, M.B., Yazici, A., Aljarrah, M., 2020. Value of convenience for taxi trips in New York City. Transport. Res. Part A Pol. Pract. 142, 85–100. https://doi.org/10.1016/j.tra.2020.10.016.
Wang, F., Ross, C.L., 2019. New potential for multimodal connection: exploring the relationship between taxi and transit in New York City (NYC). Transportation 46, 1051–1072. https://doi.org/10.1007/s11116-017-9787-x.
Wikipedia contributors, 2019. List of United States Rapid Transit Systems by Ridership. https://en.wikipedia.org/wiki/List_of_United_States_rapid_transit_systems_by_ridership. [Online. (Accessed 1 December 2019).
Wong, R., Yang, L., Szeto, W., Li, Y., Wong, S., 2020. The effects of accessible taxi service and taxi fare subsidy scheme on the elderly's willingness-to-travel. Transport Pol. 97, 129–136.
Yang, C., Morgul, E.F., Gonzales, E.J., Ozbay, K., 2014. Comparison of mode cost by time of day for nondriving airport trips to and from New York City's Pennsylvania Station. Transport. Res. Record 2449, 34–44. https://doi.org/10.3141/2449-04.
Zhang, X., Xu, Y., Tu, W., Ratti, C., 2018. Do different datasets tell the same story about urban mobility—a comparative study of public transit and taxi usage. J. Transport Geogr. 70, 78–90. https://doi.org/10.1016/j.jtrangeo.2018.05.002.
Zhao, P., Cao, Y., 2020. Commuting inequity and its determinants in Shanghai: New findings from big-data analytics. Transport Pol. 92, 20–37. https://doi.org/10.1016/j.tranpol.2020.03.006.
Zhong, H., Li, W., 2016. Rail transit investment and property values: an old tale retold. Transport Pol. 51, 33–48. https://doi.org/10.1016/j.tranpol.2018.04.010.