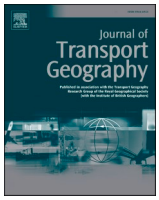
Ride-sourcing compared to its public-transit alternative using big trip data

Journal of Transport Geography 95 (2021) 103135

Contents lists available at ScienceDirect

Journal of Transport Geography

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



Ride-sourcing compared to its public-transit alternative using big trip data

Yuan Liao

Department of Space, Earth and Environment, Division of Physical Resource Theory, Chalmers University of Technology, Gothenburg, Sweden

ARTICLE INFO

Keywords:
Ride-sourcing
Public transit
Urban mobility
Travel time
Built environment
Glass-box model

ABSTRACT

Ride-sourcing risks increasing GHG emissions by replacing public transit (PT) for some trips therefore, understanding the relation of ride-sourcing to PT in urban mobility is crucial. This study explores the competition between ride-sourcing and PT through the lens of big data analysis. This research uses 4.3 million ride-sourcing trip records collected from Chengdu, China over a month, dividing these into two categories, transit-competing (48.2%) and non-transit-competing (51.8%). Here, a ride-sourcing trip is labelled transit-competing if and only if it occurs during the day and there is a PT alternative such that the walking distance associated with it is less than 800 m for access and egress alike. We construct a glass-box model to characterise the two ride-sourcing trip categories based on trip attributes and the built environment from the enriched trip data. This study provides a good overview of not only the main factors affecting the relationship between ride-sourcing and PT, but also the interactions between those factors. The built environment, as characterised by points of interest (POIs) and transit-stop density, is the most important aspect followed by travel time, number of transfers, weather, and a series of interactions between them. Competition is more likely to arise if: (1) the travel time by ride-sourcing $<$15 min or the travel time by PT is disproportionately longer than ride-sourcing; (2) the PT alternative requires multiple transfers, especially for the trips happening within the transition area between the central city and the outskirts; (3) the weather is good; (4) land use is high-density and high-diversity; (5) transit access is good, especially for the areas featuring a large number of business and much real estate. Based on the main findings, we discuss a few recommendations for transport planning and policymaking.

1. Introduction

Sustainable urban development aims to find solutions that mitigate negative environmental impacts, e.g., congestion and emissions of greenhouse gases (GHGs), brought by rapidly increasing numbers of motorised vehicles especially in developing countries. In China, for instance, the ownership of private vehicles reached 254 million in 2019, which is 9.5% more than in 2018 (National Bureau of Statistics of China, 2020). A core transportation strategy to mitigate negative environmental impacts is shared mobility, which refers to the services and resources involved in using a motor vehicle, bicycle, or other low-speed transportation mode that is shared among users, either concurrently or one after another (Shared-use Mobility Center, 2020; Shaheen et al., 2016). Public transit (PT)/mass transit and ride-sourcing (here, the latter refers to on-demand mobility services via smartphone apps to connect drivers with passengers) are both included in shared mobility.

As a mode of shared mobility in cities, ride-sourcing services become increasingly popular; one of the key questions remains unanswered: Does ride-sourcing complement, or compete with, PT? The importance of this question lies in the different GHG intensities of these two modes. About 70–80% of the variation in the GHG intensity of major passenger transportation modes can be explained by occupancy (Schäfer and Yeh, 2020). Despite both being shared-mobility modes, PT outperforms ride-sourcing on shared occupancy leading to a lower GHG intensity (IEA, 2012; California Air Resources Board, 2019). Therefore, if ride-sourcing mostly competes with PT, it may increase GHG emissions from transport systems.

The advent of trip datasets with high spatio-temporal resolution offers improved understanding of the relationship between ride-sourcing and PT in urban mobility. Traditional surveys (Rayle et al., 2016; Aarhaug and Olsen, 2018; Yan et al., 2019) are limited by their small sample sizes, deviations from actual travel behaviours, and failures to incorporate the built environment. The recent development of GPS-enabled devices allows for fast accumulation of a massive amount of spatial data, offering new opportunities. For example, the City of New York has an open data portal for taxi trips (City of New York, 2020); these data have been applied to answer a variety of questions (Qian and Ukkusuri, 2015; Kamga et al., 2015; Hochmair, 2016; Wang and Ross, 2019),

E-mail address: yuan.liao@chalmers.se.

https://doi.org/10.1016/j.jtrangeo.2021.103135
Received 10 January 2021; Received in revised form 21 June 2021; Accepted 2 July 2021
Available online 20 July 2021

including about the relationship between on-demand transport and PT (Hochmair, 2016; Wang and Ross, 2019). Wang and Ross (2019) categorise the interplay between taxi services and PT into three groups: transit-competing, transit-complementing, and transit-extending. If ride-sourcing trips tend to directly compete with PT, policymakers will need to know why, in order to mitigate negative environmental impacts, e.g., congestion and GHG emissions, from urban transportation systems.

This study attempts to explore the use of ride-sourcing trip data to understand urban mobility, in particular the relationship between ride-sourcing and PT. The explored trip attributes include travel time and the built environment of pick-up and drop-off locations. We use an open dataset made public by the largest mobile transportation platform in China. The study enriches the ride-sourcing trips with travel information by PT, hourly weather records, and Points of Interest (POIs) and transit-stop density representing the built environment of the study area. We divide the ride-sourcing trips into two categories, transit-competing (48.2%) and non-transit-competing trips (51.8%). A ride-sourcing trip is labelled transit-competing if and only if it occurs during the day (from 6 am to 11 pm) and there is a PT alternative such that the walking distance associated with it is less than 800 m for access and egress alike. By comparing the two categories of ride-sourcing trips using a glass-box model, the enhanced Generalised Additive Model (eGAM), this study reveals how the trip attributes and built environment are linked to the competition between ride-sourcing and PT, and the applications for transport planning.

The remainder of this paper is organised as follows. Section 2 reviews the related work and proposes study objectives given the gaps in the literature. Section 3 describes the applied ride-sourcing trips and built environment characterisation of the study area, and the model construction. The descriptive analysis and model results are presented in Section 4, and Section 5 discusses the findings and considers future research. Section 6 concludes the study.

2. Literature review

2.1. Roles of ride-sourcing and public transit

Ride-sourcing is an emerging mode of shared mobility, with rapid growth worldwide in the use of phone-based ride-hailing applications such as Uber, Lyft, and DiDi Chuxing. Ride-sourcing provides flexible on-demand services. In contrast, PT offers scheduled services in rigid networks where travellers need to adapt their plans accordingly (Winter et al., 2018). On-demand transport services (e.g., ride-sourcing) are found to be a significant transport mode in urban areas and complements PT through its door-to-door transportation without transfers (Hochmair, 2016). Regarding whether ride-sourcing absorbs the demand for travel by PT, a study in Amsterdam found most PT trips are replaced by active modes (e.g., bike) and most car trips by ride-sourcing (Narayan et al., 2019). Becker et al. (2020) found that the impact of ride-hailing appears less positive; seemingly, the clear efficiency has been detected when substituting for PT in lower-density areas. The interplay between ride-sourcing and PT is complex and environment-dependent therefore, regarding if and when ride-sourcing competes with PT, different conclusions exist in the literature.

Despite the elusive relationship between ride-sourcing and PT, the narrative of shared mobility sometimes depicts PT as marginalised by new ride-sourcing solutions such as Uber and Lyft, due to tight government PT budgets (Brustein, 2016). However, the environmental impact of ride-sourcing as a complementary or alternative solution to PT remains a major concern. PT is considered environmentally friendly and supportive of urban liveability and has the highest vehicle occupancy and greatest capacity (Currie, 2018). PT is estimated to emit 4–22 g of CO_2 per passenger-mile travelled (gCO_2/PMT) depending on the mode (IEA, 2012). As for ride-hailing, the California Air Resources Board estimates that vehicle fleets run by ride-hailing services such as Uber and Lyft emit 301 gCO_2/PMT, approximately 50% more than the state-wide passenger vehicle fleet average of 203 gCO_2/PMT (California Air Resources Board, 2019). Therefore, PT outperforms both private car use and ride-hailing services in terms of carbon emissions.

2.2. Exploring factors of mode choice

Based on empirical trip data, travel time (e.g., Schafer and Victor, 2000), built environment (e.g., Ewing and Cervero, 2010), and socio-demographics (e.g., Shirgaokar, 2018) are considered the most critical factors. For example, Wang and Ross (2019) found that 59.5% of taxi trips serve disabled, low-income, elderly, retired, or unemployed people. A meta-analysis found that PT use is related to proximity to transit and street networks and mixed land-use (Schafer and Victor, 2000). Utility-based decision models have widely incorporated two essential travel-mode attributes, travel time and cost (De Vos et al., 2016). A reduction of travel time is known to encourage more people to shift from private car to PT (Redman et al., 2013).

Data collection has widely relied on surveys to better understand the impact of increased availability of ride-sourcing. Rayle et al. (2016) described the findings from a survey of 380 ride-sourcing users where half of the trips replaced modes other than taxi service, e.g., PT. An expert survey based on 76 respondents was used to project how ride-sourcing and automated vehicles affect on-demand transport markets (Aarhaug and Olsen, 2018). Based on the 2015 US census data, Reck and Axhausen (2019) calculate travel times by ride-sourcing and PT with Google Directions API and explore the potential for ride-sourcing to solve first/last-mile issues associated with PT. A web-based survey with 4473 respondents, including university faculty and staff members and students, was used to model impacts on ridership of integrating ride-sourcing with public transit (Yan et al., 2019). That study concluded that ride-sourcing complements public transit by enhancing last-mile transit access (Yan et al., 2019). Despite having abundant information on traveller socio-demographics, surveys have been constrained by small sample sizes for traditional ones, sometimes biased samples for web-based ones using convenient sampling, and deviations from actual travel behaviours due to hypothetical bias (Murphy et al., 2005). Moreover, the lack of precise geolocation prohibits relating results to the built environment.

2.3. Emerging big trip data

Given limited resources, the still elusive relationship between ride-sourcing and PT is of great relevance for policymaking. The recent development of GPS-enabled devices accumulates a massive amount of spatial data to exploit the travel demands, bringing new opportunities. Studies have used large amounts of real-world data to analyse ride-sourcing, ride-hailing, and PT in urban mobility. Qian and Ukkusuri (2015) use large-scale real-world data to model taxi demand in New York City. With the same dataset, 147 million taxi-trip records covering 10 months, Kamga et al. (2015) attempt to reveal the impact of time and weather on taxi ridership. Welch et al. (2020) apply big data analytic tools to explore the factors that motivate massive amounts of trips by transit, taxi, and bike-sharing in Washington, D.C. Wang and Ross (2019) discuss the relationship between taxi and transit in three categories: transit-competing, transit-complementing, and transit-extending. Empirical analysis evidenced that PT oftentimes can be replaced by taxi service when PT access is good: Wang and Ross (2019) found that 58.5% of taxi trips in New York City have both pick-up and drop-off location within a 2-mile radius of a transit station during the times when transit services are available. If ride-sourcing trips tend to directly compete with PT as such, policymakers should address the reasons why riders choose ride-sourcing over PT (Welch et al., 2020).

To summarise, previous studies aim to understand the roles of different modes while their interplay is rarely discussed. Few studies directly tackle the relationship between ride-sourcing and PT using big trip data, not to mention the case of developing countries. Therefore,

this study uses big data analysis to explore urban mobility by ride-sourcing, at a trip-level comparison with the PT alternative. We explore the travel demand between PT and ride-sourcing, particularly the tendency of ride-sourcing to replace PT trips.

2.4. Methods

As the literature shows, more and more geolocated data are made freely available to the public. However, coverage of a large area and population is often achieved at the cost of rich detail, including, for example, trip purpose, compared to traditional survey-based data. Open data often only contain the geolocation of origin/destination and partial trajectories without trip purpose. To make full use of the data, they have to be "enriched", i.e., incomplete data have to be supplemented, often using external data sources (Allen and Cervo, 2015). In order to reveal the shared-use mobility competition at the trip level, Welch et al. (2020) combine data from various sources, including taxi trips, metro line trips, census, and OpenStreetMap. OpenStreetMap provides crowd-sourced built-environment characteristics and transportation-network connectivity for a better explanation of the observed trip patterns.

Besides information from external data sources, advanced techniques are applied in order to leverage more and more big trip data. Given the importance of the built environment in understanding the interplay between modes, the application of unsupervised learning in urban land use provides a new angle on incorporating the built environment in data enrichment, contributing to the analysis of the spatial distribution of trips. Typically, the built environment can be quantified by population and employment density, road network characteristics, land-use diversity/entropy, population and employment diversity/entropy, and accessibility indicators (Yu and Peng, 2019). These aspects are treated as independent features directly explaining observed demand for travel by ride-sourcing or PT; however, they are in fact not independent. Moreover, this high dimensionality makes it challenging to gain useful insights. Along with the increased availability of big geodata, studies have started using data on points of interest (POIs) to cluster space into functional urban regions (Gao et al., 2017; Hu et al., 2020). These efforts attempt to understand urban space comprehensively, which benefits the interpretation of the spatial distribution of travel demand derived from trip records.

Another key technique, community detection, has been widely used with large amounts of mobility traces to uncover the interactions between locations, especially their underlying structure, e.g., which locations tend to be connected (Sobolevsky et al., 2014). For taxi-trip records, for example, understanding which locations are closely connected by trip origin and destination benefits taxi-fleet supply management. For ride-sourcing services, this network structure informs drivers seeking passengers and therefore increases average occupancy. Liu et al. (2015) reveal the city structure of Shanghai using community detection based on a large amount of taxi-trip records, which benefits transportation planning in general; however, the policy implications are not explicitly explored. In another study, Zhang et al. (2018) compare the spatial structure of taxi and transit trips in Singapore and reveal their different roles in connecting certain places in the city. However, the potential for understanding the relationship between different modes is not fully exploited.

After enrichment, big trip data feature a large volume and high dimensionality. For trip-based mode choice and classification of trips, common methods include generalised linear models (McCullagh, 2018), such as multinominal logit (e.g., Welch et al., 2020) and binary logit model (e.g., Wang and Ross, 2019). These methods are widely used due to the simple form, intelligibility, and potential for scenario simulations. On the other hand, increasing attention has been paid to applying machine-learning techniques, however, using them comes with challenges in interpretability: Insights about the data and the task the machine solves are hidden in increasingly complex models (ch1.2, Molnar, 2020). Recent advances in glass-box models aid interpretation and are therefore beneficial in the use of machine learning for exploring the relationship between transit and ride-sourcing. Common glass-box models include linear regression models and their extensions, logistic regression models, and decision trees. These glass-box models hold the potential to better synthesise trip dimensions. For instance, in the generalised additive model (GAM) (Hastie and Tibshirani, 1990), a generalised linear model, the linear part of the variable depends linearly on unknown smooth functions of the independent variables, and the model construction focuses on the inferences about these smooth functions. The recent machine-learning techniques have enhanced the traditional GAM by bagging, gradient boosting, and automatic interaction detection (Nori et al., 2019). Compared with classic glass-box models such as logit models, this enhanced GAM generally delivers more accurate results, while keeping them insightful and easy to visualise. Therefore, we use this enhanced version of GAM in the present study.

2.5. Study objectives

Though it is important to understand the interplay between ride-sourcing and PT, the potential for ride-sourcing services to replace PT trips has largely been overlooked (Wang and Ross, 2019; Narayan et al., 2019; Welch et al., 2020). The relationship between ride-sourcing and PT remains elusive, especially in developing countries where data are often lacking. Meanwhile, a rapidly growing body of literature uses advanced techniques in machine learning and network science with big trip data to model urban structure. However, they remain under-exploited on revealing the impacts of trip attributes and the built environment on the relationship between ride-sourcing and PT.

This study explores the competition between ride-sourcing and PT through the lens of big data analysis. For the characterisation of trip attributes and built environment, we incorporate functional urban regions identified using POIs with clustering analysis, transit access, and the community structure of ride-sourcing demand. We use a glass-box model that predicts whether a ride-sourcing trip directly competes with its alternative PT, providing intelligible outputs that can easily be visualised. The main factors include travel time for ride-sourcing and PT, weather condition, functional regions, transit access, and demand-based communities of pick-up and drop-off zones. Specifically, three nested questions are explored, as shown below:

- Does ride-sourcing compete with public transit?
- What trip attributes and built environment are linked to the competition?
- What are the implications for policymaking?

3. Methodology

To compare ride-sourcing with its PT alternative, we use a set of big trip data collected from the largest ride-sourcing platform in China, as described in Section 3.1.

The methodological framework is illustrated in Fig. 1. In preprocessing the original dataset (Section 3.1.1), we first filter out abnormal request records and enrich each record with the travel information for its PT alternative assuming the same departure time, origin, and destination as well as with the weather information for the departure time. Moreover, we detect the community structure of the ride-sourcing origin-destination matrix created by connecting all the pick-up and drop-off zones. By doing so, we divide the study area into sub-regions based on the ride-sourcing travel demand. These demand-based sub-regions help us better identify the trend for the competition between ride-sourcing and PT.

The original dataset consists of a series of records with the origins and destinations of ride-sourcing trips but without any informative environmental context. In order to know more about the built environment of the pick-up and drop-off spots, we identify the functional

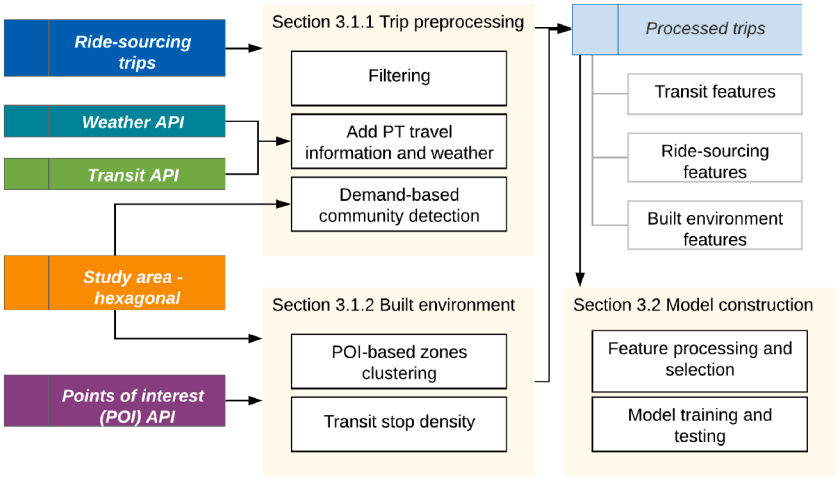


Fig. 1. Methodological framework. The arrows mark the flow of data.

clusters using points of interest (POIs) of the zones in the study area and quantify the transit access density in the zones (Section 3.1.2).

In the model construction (Section 3.2), we first label the processed ride-sourcing trips as transit-competing or non-transit-competing based on the time of day and the walking distance to and from transit stations. Next, we process the features to eliminate multicollinearity and select the qualified features. Finally, we construct a model to characterise the two categories in terms of the trip attributes and the built environment in order to answer the proposed research questions.

3.1. Data description and processing

The original dataset used in this study is a complete sample of the ride request data registered in Chengdu, China from November 1st to November 30th, 2016, provided by DiDi Chuxing GAIA Open Dataset Initiative (DiDi Chuxing, 2020b). Didi Chuxing is a mobile transportation platform covering ride-sourcing among other services (DiDi Chuxing, 2020a). It provides over 10 billion passenger trips a year. The ride requests included contain order ID, start and end time, and GPS coordinates of pick-up and drop-off locations. No individual information or trip purpose is available from this dataset. There are around 7.1 million ride requests recorded within Chengdu City during 22 weekdays in November 2016.

Chengdu is the sixth largest city by urban-area population in China, the capital of Sichuan province in southwestern China. The population is 16.6 million, area 25,248 km^2, and the gross domestic product (GDP) per capita 14,600 $/year (Chengdu Bureau of Statistics, 2020). Chengdu has a 24-h PT system, although service is limited from 11 pm to 6 am. As shown in Fig. 2 where the bus/metro stations are from Baidu Place Application Programming Interfaces (APIs) (Baidu Maps, 2020a), the study area is divided into hexagonal cells, each of which has a short diagonal of 500 m. The hexagonal sampling grid is selected because it allows better distribution of the centroids of zones as sampling points, and it requires fewer grid cells compared to a similar grid of squares (Burdziej, 2019). There are 8279 hexagonal zones that have at least one location that is either the origin or destination of a ride request.

3.1.1. Preprocessing ride-sourcing trips

This study focuses on a number of weekdays randomly selected due to limited access to retrieving transit travel information from commercial APIs. The ride-sourcing trip records contain some abnormal travel times due to logging errors. According to the travel time distribution of the trips, we identify the outlier values as being in the top or bottom ten-thousandth of a percent. Considering that longest possible trip (120 km) takes no more than 200 min with a low speed of 35 km/h, we keep the

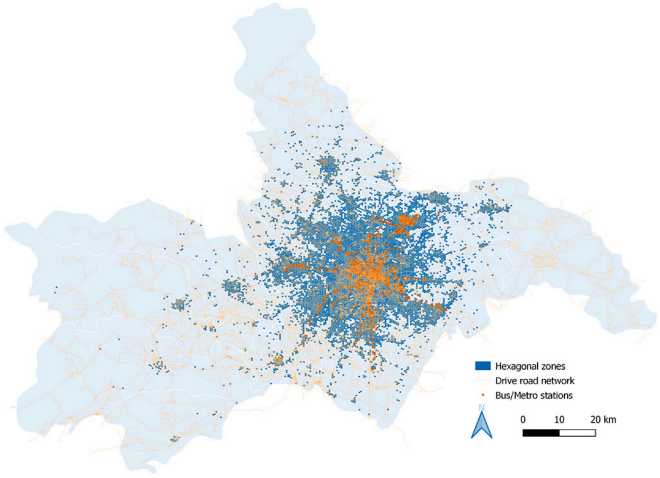


Fig. 2. Chengdu. The study area includes Chengdu City and the surrounding cities10

trips that have travel times within the range of 5 min–200 min. In total, there are 4.27 million trip records used in this study.

For these ride-sourcing trips, PT is an alternative to get to the same drop-off location from the pick-up location. In order to get the travel details of taking PT, we feed the pick-up time and the pick-up and drop-off locations of each ride-sourcing record into Baidu Transit API (Baidu Maps, 2020b). This step assigns to each ride-sourcing trip record the trip information for taking PT, including travel time by PT, walking distance to and from transit stations, and the number of boardings. We also enrich the trips with the hourly records of weather in Chengdu during November 2016.[1]

The occurrence of ride-sourcing trips is not evenly distributed in space; therefore, dividing the study area into sub-regions based on travel demand helps in better understanding the spatial patterns. The ride-sourcing travel demand can be represented by the complex network connecting all the pairs of pick-up and drop-off zones. In order to divide the study area based on the ride-sourcing travel demand, we detect the community structure of this complex network. We apply Combo, an algorithm that iterates over a sequence of moves that alter the community structure of the network to maximise the modularity gain and automatically decides the optimal number of communities (Sobolevsky

[1] Retrieved from OpenWeatherMap API at https://openweather.co.uk/.

et al., 2014). Combo algorithm has been shown to fit particularly well for spatial networks (Sobolevsky et al., 2014), such as taxi trip records (Huang et al., 2018). As a result of demand-based community detection, the zones within a given community have a higher likelihood of connecting to each other than to zones in other communities (Barabási et al., 2016, p. 322). These communities of pick-up and drop-off zones are added to each ride-sourcing trip record, specifying between which sub-regions the trips were generated.

3.1.2. Characterising the built environment

Built environment refers to human-made environment which is a multidimensional concept that incorporates roads, land uses, buildings, etc.; it can be measured with their design, density, and diversity (Cervero and Kockelman, 1997). Urban functional regions describe urban land focusing on land-use types and human activities (Hu et al., 2020), representing the dynamic manifestation of the built environment where Points of Interest (POIs) have been widely used (e.g., Gao et al., 2017; Hu et al., 2020). In addition to the dynamic aspect of the built environment, transit access has been used to describe the built environment (Cordera et al., 2017). As broad as the concept of the built environment is, there are many relevant measures applied in the literature. However, given data availability and the focus on the dynamic aspect, this study mainly uses POIs and transit access (transit-stop density) for characterising the built environment of the study area.

We create functional clusters using POIs as a characterisation of the built environment in the study area. The place API in Baidu Maps (Baidu Maps, 2020a) is used to retrieve POI data in each zone with a search radius centring the zone's centroid. Given that the short diagonal of the hexagonal cell is 500 m, a search radius of 500 m is selected to cover the POIs of the whole study area and duplicated POIs are removed afterwards. Under the definition of the POI API (Baidu Maps, 2020a), there are 18 types of POIs used: food, hotel, shopping, life services, beauty, tourism, leisure, sports, education, culture, medical services, automobile services, finance, real estate (office buildings, residential areas, dormitories), industrial zone, governmental agencies and organisations, access (exits or entrances to highways, parking lots, etc.), and natural attractions.

After retrieving POIs, each zone is represented by a vector of 18 POI type counts. Principle Component Analysis (PCA) is applied to the Min-Max normalised samples to keep 95% of the variance. The dimension-reduced samples are fed to a K-means clustering process to cluster the 8279 zones in the study area. Originated in the 1950s, the K-means algorithm clusters samples so that the squared error between the empirical mean of a cluster and the vectors in the cluster is minimised (Jain, 2010). K is selected by trying a variety of values from 2 to 20 to maximise the silhouette value, which quantifies how the samples are appropriately clustered (Rousseeuw, 1987; Gao et al., 2017), where samples within a cluster are similar to each other and samples that belong to different clusters are very different. This step forms functional regions based on the POI profile of the zones.

Besides POI-based functional clusters of the study area, we introduce another zonal indicator: transit-stop density (number of stops per km^2), which measures the transit supply of a given spatial zone (Barajas and Brown, 2021). These functional clusters and transit-stop density of pick-up and drop-off zones are added to each ride-sourcing trip record, revealing the built environment of where the trips originated and of their destinations.

3.2. Model construction

3.2.1. Defining transit-competing trips

In order to study the relationship between ride-sourcing trips and PT as their alternative, we need to first define transit-competing ride-sourcing trips. Studies have shown that the willingness to walk is a major factor constraining PT use (Tolley, 2016). Walking for 5 min or a distance of 477 m to a transit station can be assumed as a measure of accessibility in a European city context (Sarker et al., 2019). Some studies use 800 m when assessing the performance of transit systems (e.g., Ryan and Frank, 2009). A study on bus rapid transit in the context of Chinese cities found that 80% of survey respondents reported an access/egress walking distance of less than 800 m (Jiang et al., 2012).

Therefore, this study defines transit-competing ride-sourcing trips as follows: If a ride-sourcing trip, when instead served by PT, were to have had an access/egress walking distance of less than 800 m (each), and depart between 6 am–11 pm, it is called a transit-competing trip ($y = 1$), otherwise it is non-transit-competing ($y = 0$).

After labelling the trip records as transit-competing or non-transit-competing, we select features to model to distinguish them. The model reveals the relationship between ride-sourcing and transit regarding the trip attributes and the built environment.

3.2.2. Feature processing and selection

To characterise whether a ride-sourcing trip is transit-competing, a series of candidate trip variables are created from the processed data (Section 3.1), as summarised in Table 1.

Among the candidate explanatory variables, we expect more than two variables to be highly correlated, such as trip distance and travel distance. This multicollinearity issue may result in model overfitting (Dormann et al., 2013). To detect multicollinearity, the variance inflation factor (VIF) analysis (James et al., 2013) is applied to evaluate all the candidate variables. We remove the variables with a VIF above 10 (Mason et al., 2003): trip distance, travel distance by ride-sourcing, cost of ride-sourcing, and travel distance by PT.

3.2.3. Enhanced generalised additive model (eGAM)

The enhanced generalised additive model (eGAM) is selected to predict whether a ride-sourcing trip is transit-competing by letting x_i range over the selected variables in Table 1 and the top eight interactions between them. eGAM originates from the traditional Generalised Additive Model (GAM) (Hastie and Tibshirani, 1990):

$$g(E[y]) = \beta_0 + \sum f_i(x_i) \tag{1}$$

where g is a link function connecting the expected value of y with the right part of the equation, β_0 is a constant, and $f_i(x_i)$ is an unknown smooth function of x_i. The logit function is a common link function for binary classification. GAM has subsequently been modified into a model called GA^2M (Lou et al., 2013) that allows interactions between explanatory variables to be captured:

$$g(E[y]) = \beta_0 + \sum f_i(x_i) + \sum f_{ij}(x_i, x_j) \tag{2}$$

This increases accuracy while keeping a high level of intelligibility. The training process of GA^2M finds the form of variable smooth functions. GA^2M is further enhanced by modern-machine learning techniques to train GA^2M faster while allowing for large datasets (Nori et al., 2019). It also enables automatic interaction detection; therefore, we use the label eGAM in this study.

eGAM is applied with a randomly-selected 75% of trip records for training and the rest for testing. For each given record sample, every feature or feature interaction returns a score i.e., $f_i(x_i)$ and $f_{i,j}(x_i, x_j)$. Whether this sample is predicted as transit-competing is dependent on the summation of those scores. Therefore, a single feature or feature interaction scoring above 0 increases the chance of the sample being transit-competing ($y = 1$).

4. Results

In this section, we first describe the basic statistics, pickup and drop-off hot spots, demand-based communities, and built environment of the study area based on the selected 4.27 million trip records (Section 4.1). Next, the model results are presented, covering model performance and

Table 1
Candidate explanatory variables. Variables in bold are the ones fed into the model after the feature selection.

Variable type	Variable	Unit	Description
Ride-sourcing	Trip distance	km	Straight-line distance between pick-up and drop-off locations
	Travel distance by ride-sourcing	km	Network distance between pick-up and drop-off locations by driving
	TT by ride-sourcing	min	Time duration of a trip record
	Cost of ride-sourcing	RMB	Estimated based on the taxi fee
	Weather	–	Clouds, Clear, Haze, Fog, Mist, Rain
	Demand community (pick-up zone)	–	Pick-up location community
	Demand community (drop-off zone)	–	Drop-off location community
PT	Travel distance by PT	km	Network distance between pick-up and drop-off locations by PT
	TT ratio excl. access/egress walking	min	Travel time by PT excluding access/egress walking divided by TT by ride-sourcing
	# of boardings	–	Number of boardings for taking transit
Built environment	**Functional cluster (pick-up zone)**	–	Pick-up location cluster
	Functional cluster (drop-off zone)	–	Drop-off location cluster
	Transit-stop density (pick-up zone)	$1/km^2$	Transit-stop density of pick-up location
	Transit-stop density (drop-off zone)	$1/km^2$	Transit-stop density of drop-off location

the scores of single features and feature interactions for transit-competing and non-transit-competing ride-sourcing trips (Section 4.2). Finally, Section 4.3 provides a simplified summary of the situations with above-average likelihood that a trip will be transit-competing, including detailed discussion of two specific cases.

4.1. Descriptive analysis

4.1.1. Ride-sourcing trips

In the 4.27 million ride-sourcing trips analysed, the by-definition transit-competing trips (i.e., those generated between 6 am and 11 pm that had a PT alternative for which both access and egress walking distance was less than than 800 m) account for 48.2%. The three criteria for transit-competing and non-transit-competing trips (time of day and access and egress distances) are illustrated in Fig. 3. During the day, the number of transit-competing trips is twice that for the non-transit-competing trips, while the gap starts to decrease at 7 pm (Fig. 3A). Therefore, the ride-sourcing trips that people take at night start to become non-transit-competing as compared with in the daytime. As for the walking distances (Fig. 3B), 48% of the ride-sourcing trips have a PT alternative that requires either access or egress walking distance of more than 800 m.

The ride-sourcing trips have their spatial patterns as revealed by the demand-based communities. They divide the study area into three sub-regions (Fig. 4A), the North, the South-West, and the South-East, within which the zones are closely connected by the ride-sourcing trips. Fig. 4B shows the number of trips between and within the three communities; of these nine types of trips, the largest single category is trips within the South-West. Moreover, the connections between the South-West and the other two are also busier than the other connections.

Besides the heterogeneous overall spatial distribution of the ride-sourcing trips, we also find distinct spatial patterns between transit-competing trips and non-transit-competing trips. Their statistically significant hot spots (cells) are shown in Fig. 5 where all the cells shown are statistically significant detected by Getis-Ord Gi* (Getis and Ord, 2010) with Z-score $\geq$1.96 and $p < 0.05$. These are hot spot cells at the 95% confidence level. The group of cells on the southwestern side are Chengdu Shuangliu International Airport. The group of cells on the southeastern side are a railway station. The non-transit-competing trips tend to have a more spread-out distribution of pick-up and drop-off hot spots, including the international airport, the railway station, and the northern area. On the other hand, the transit-competing trips tend to concentrate in the central area, where the railway station appears to be the drop-off hot spot but not the pick-up hot spot. This implies that when the destination is the railway station, ride-sourcing is more competitive with PT, compared to when the railway station is the origin.

4.1.2. Built environment

The built environment of the pick-up and drop-off spots are characterised by POIs in the zones in the study area. Seven functional clusters are created, see Fig. 6A-B. Cluster Centre features the centre of Chengdu City and the centres of the surrounding small cities; it has the highest number of almost all the POIs except for automobile services. Cluster Outer-residential is located around Chengdu City, although at a distance from the central area; it features a large number of residences and a smaller number of other POIs. Clusters Centre-business and Transition are located between Clusters Centre and Outer-residential in a ring structure; they have a similar structure of POIs as Cluster Centre, but Cluster Centre-business, bordering the centre of Chengdu City, also features many businesses.

The surrounding cities have the same ring structure but the order of clusters is Centre, Centre-business, and Residential-business from the inner circle to the outer, as illustrated in Fig. 6B. Compared with Chengdu centre, these surrounding cities are less developed and therefore have simpler land-use structure. Cluster Residential-business features a roughly equal number of residences and businesses, while the

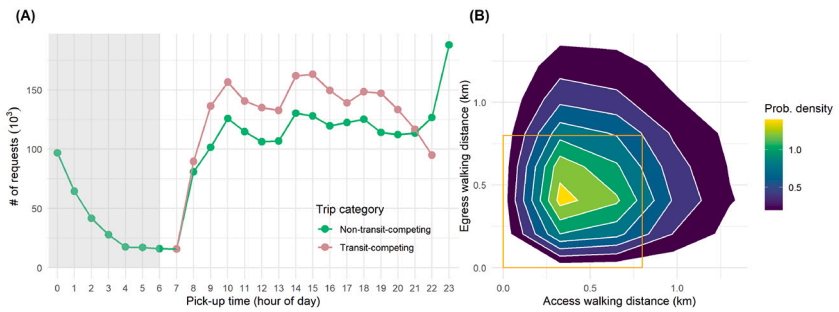


Fig. 3. Trips by time of day and walking distance. (A) Temporal distribution of the pick-up time of requests. Shaded area shows the indicator range of being non-transit-competing. (B) Probability density of access and egress walking distance. The area in the rectangle shows the indicator range of being transit-competing.

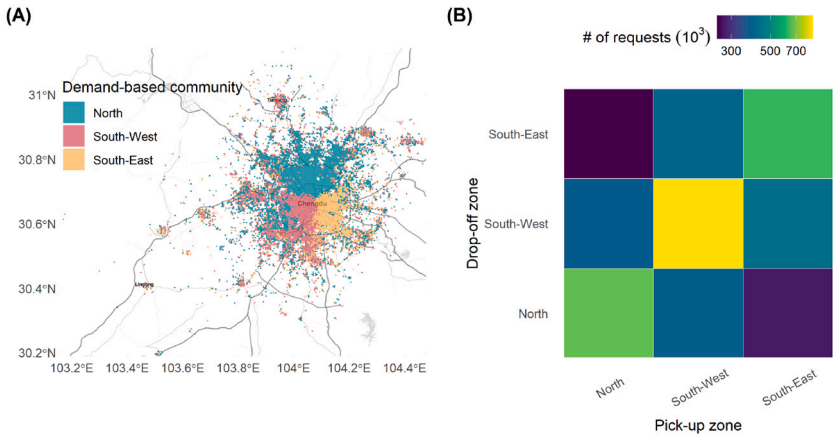


Fig. 4. Communities. (A) Spatial distribution. (B) Number of trips within each community or between communities.

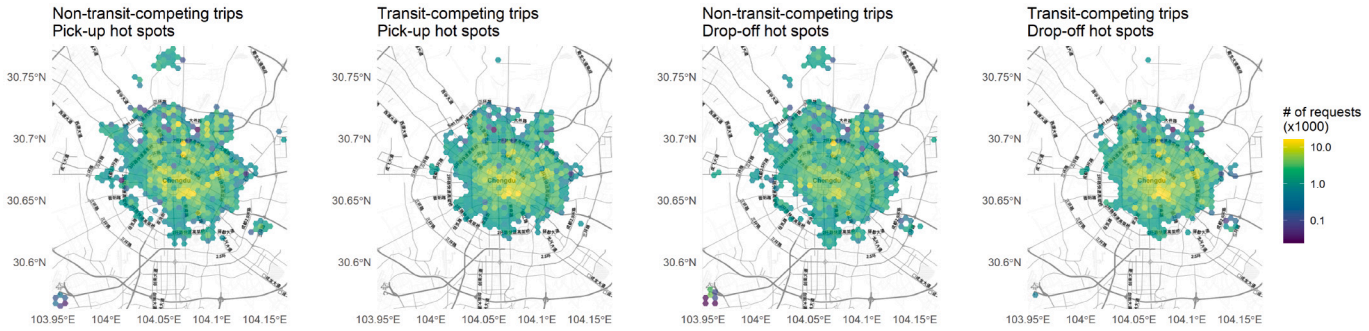


Fig. 5. Hot spots of ride-sourcing trips.

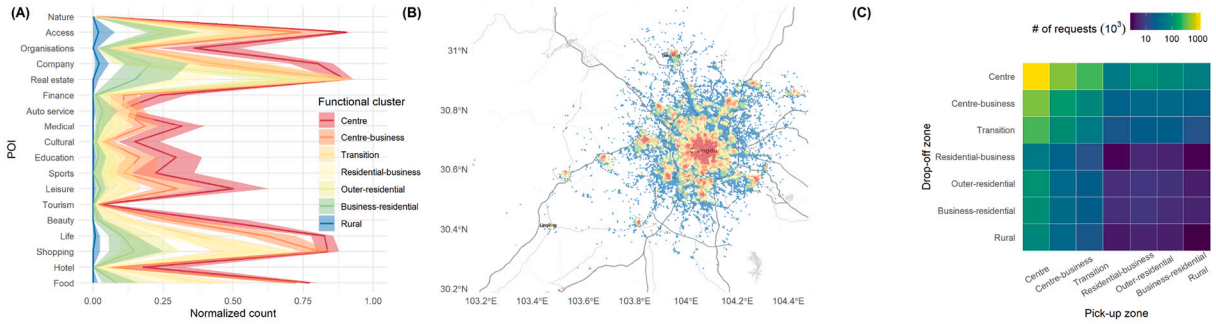


Fig. 6. Functional clusters of zones. (A) The normalised number of POIs (of 18) in the functional clusters of zones. The shaded area indicates the range from 25th percentile to 75th percentile. (B) Spatial distribution. (C) Trip count between and within the zones of different functional clusters.

number of other POIs is close to that in Cluster Outer-residential. Cluster Rural is located on the perimeter of the study area, where there are only a small number of businesses and access points to major roads (access). In general, Cluster Business-residential borders on Cluster Rural, while being closer to the city centre; there are more places related to living (food, shopping, and life), especially more residences (real estate), and greater road density (access).

To summarise the observed patterns (Fig. 6A), Cluster Centre is the first tier with the highest land-use intensity and diversity, which generates and attracts the most ride-sourcing trips (Fig. 6C). As the second tier, Clusters Centre-business, Transition, and Residential-business have a moderate level of land use, followed by Clusters Outer-residential, Business-residential, and Rural, as the third tier, in the transition area between the main city and the surrounding area. To take a closer look at these clusters, we define the share of commercial POIs per zone as the share of POIs of finance, beauty, life, shopping, hotel, and food, given these are POIs for the provision of goods or services. The share of commercial POIs of the clusters are Residential-business (52%), Centre-business (47%), Centre (44%), Transition (41%), Business-residential (31%), Outer-residential (29%), and Rural (21%) in descending order.

The descriptive statistics based on the processed trip attributes and the built environment of the ride-sourcing trips used for modelling are summarised in Table 2.

4.2. Model results

4.2.1. Model performance and feature importance

Using the selected features described in Table 2, the eGAM model performs well, as quantified by Area Under the ROC Curve (AUC = 0.70), a indicator ranging from 0 to 1 (the higher, the better a model performs).

Fig. 7 illustrates the importance of single features and the detected

Table 2
Descriptive statistics of the ride-sourcing trips used for modelling. Mean for continuous variables and % for categorical variables.

Variable	Levels	Mean (SD) or %
TT by ride-sourcing, min	–	22.30 (12.81)
TT ratio excl. access/egress walking	–	1.81 (0.85)
Transit-stop density (pick-up zone), 1/km^2	–	12.38 (12.43)
Transit-stop density (drop-off zone), 1/km^2	–	12.66 (12.74)
Weather	Clear	0.9
	Clouds	42.2
	Haze	4.8
	Fog	1.4
	Mist	41.8
	Rain	8.9
Demand community (pick-up zone)	North	30.2
	South-West	38.7
	South-East	31.1
Demand community (drop-off zone)	North	30.8
	South-West	39.2
	South-East	30.0
# of boardings	1	57.7
	2	35.6
	3	6.0
	4	0.6
	5	0.04
	6	0.003
Functional cluster (pick-up zone)	Centre	55.2
	Centre-business	18.7
	Transition	11.8
	Residential-business	2.9
	Outer-residential	4.5
	Business-residential	4.0
	Rural	3.0
Functional cluster (drop-off zone)	Centre	54.9
	Centre-business	18.0
	Transition	12.1
	Residential-business	2.7
	Outer-residential	4.4
	Business-residential	4.3
	Rural	3.6

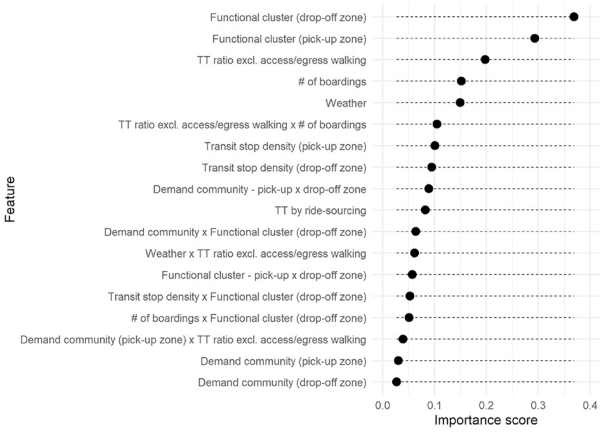


Fig. 7. Feature/feature interaction score indicating importance in predicting whether a ride-sourcing trip is transit-competing.

interactions between some of the features. The most important features that determine whether a trip is transit-competing are the functional clusters of the ride-sourcing trip's drop-off and pick-up zones (Centre, Centre-business, Transition, Residential-business, Outer-residential, Business-residential, or Rural). The TT ratio excluding access/egress walking is the third most important feature, measuring the travel time disparity between ride-sourcing and its PT alternative, followed by # of boardings and weather. The interactions between demand-based communities are important, but not demand-based community alone since they rank the least important. In addition, a few other detected feature interactions play an important role such as the interactions between # of boardings, TT ratio, functional cluster, weather, and transit-stop density.

4.2.2. *Two categories of ride-sourcing trips*

The categorisation of a ride-sourcing trip as transit-competing is affected by both the feature components, $f_i(x_i)$, and the components of the interactions between them, $f_{i,j}(x_{i,j})$, but it is the summation of all the component scores that determines the prediction outcome for a trip's category. Therefore, a score above zero means a tendency to be transit-competing ($y = 1$), while $y = 0$ for the score below zero.

Fig. 8 shows the impact of single features on the tendency of a ride-sourcing trip to be transit-competing. Such a tendency decreases consistently when TT by ride-sourcing increases (Fig. 8A). Most ride-sourcing trips have a travel time below 15 min, where they tend to be transit-competing.

The greater the TT ratio excluding access/egress walking (Fig. 8B), the longer it takes for the PT alternative relative to ride-sourcing. Of ride-sourcing trips, 50% have a TT ratio below 1.5; they tend to be non-transit competing. When the TT ratio increases, the competition tendency increases. However, when the TT ratio further increases above 3.5, the tendency of being transit-competing decreases again.

Fig. 8C suggests that the more boardings, the greater the probability of a trip competing with its transit alternative. This suggests that despite short access and egress walking distances for those transit-competing trips, the competition is likely to happen if the number of transfers between origin and destination is large.

The weather also affects whether a ride-sourcing trip has a feasible PT alternative (Fig. 8D); trips generated under in fog, mist, or rain are less likely to be transit-competing as compared with the other weather conditions.

Regarding the effect of the built environment (Fig. 8E and F), the ride-sourcing trips are more likely to be transit-competing when they have pick-up or drop-off locations in Cluster Centre. Outer-residential and Rural, which are located at the outer ring of the study area, are the areas associated with non-transit-competing trips.

The transit-top density also affects the tendency of a trip to be transit-competing (Fig. 8G and H); when the transit-stop density is above 13 (1/km^2), the competition is more likely to happen. On the other hand, when such a density is low, these ride-sourcing trips tend to fill the demand gap where transit access is insufficient.

Fig. 9 shows the impact of pairwise interactions on the tendency of a ride-sourcing trip to be transit-competing. The TT ratio interacts with # of boardings, weather, and the functional cluster of the drop-off zone (Fig. 9A–C). When the PT alternative requires one transfer and the on-board time is more than twice the travel time by ride-sourcing (Fig. 9A), the trips tend to be transit-competing. However, when the # of boardings is greater than 2, the competition tends to happen even for those trips that do not require much longer time by taking PT. Fig. 9B suggests that the non-transit-competing trips originated from the South-West area features both by-definition lengthy access/egress walking and the great disparity between travel time by PT and by ride-sourcing. Weather and TT ratio display an interesting interaction pattern (Fig. 9C). When the weather is good (clear), disproportionally long travel time by PT (large TT ratio) increases the tendency of a trip being transit-competing. However, when the weather is not ideal (rain), the trips with relatively short travel time by PT (small TT ratio) are more

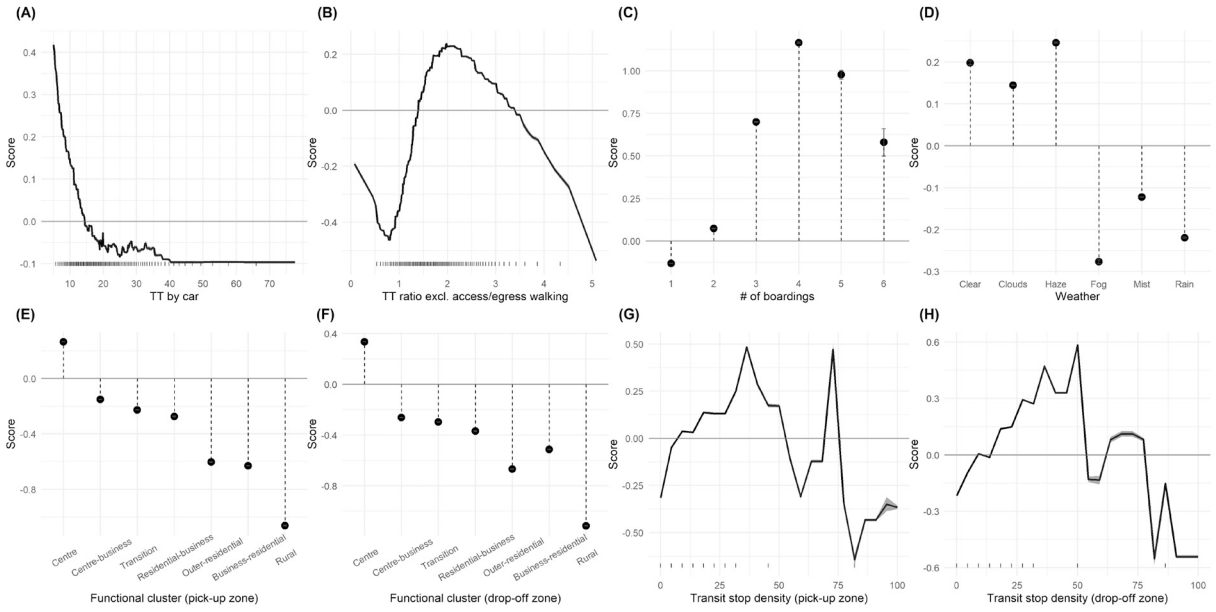


Fig. 8. Feature components for the model trained on the 3.2 million ride-sourcing trip records. Black vertical lines indicate the value of ith percentile ($i = 1, 2, \ldots, 99$). Error bars show the 95-percentile confidence level of the score curve.

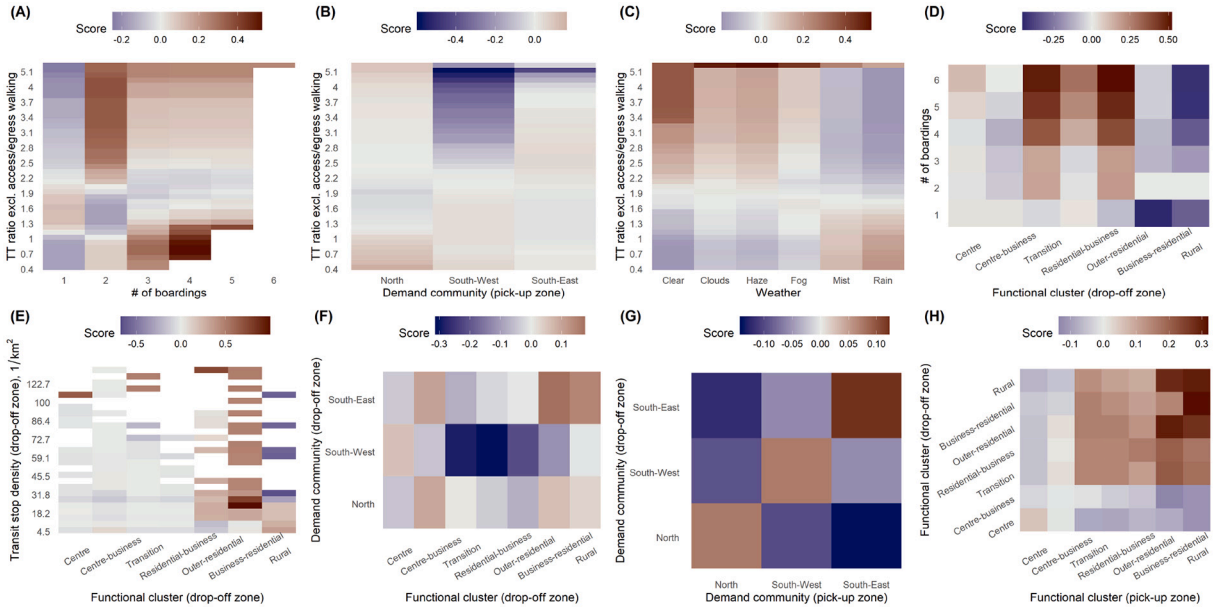


Fig. 9. Heat map of the score for the pairwise interaction components in the model, $f_{i,j}(x_{i,j})$. The blank areas have fewer than five ride-sourcing trip records, so the probability score is assumed to be unreliable. The areas coloured red and blue increase and decrease the probability of being transit-competing, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

likely to be transit-competing.

The variable # of boardings interacts with the built environment of the drop-off zone (Fig. 9D): ride-sourcing trips tend to compete with PT when they have a drop-off location in Clusters Residential-business, Transition, and Outer-residential, and the PT option has multiple transfers.

Fig. 9E suggests that the trips attracted to the zones of moderate to high transit-stop density in Cluster Business-residential are more likely to be transit-competing despite good transit access in these areas.

Fig. 9F shows the effect of the interactions between the demand-based communities and the functional clusters regarding where the trips were headed. If the trips have a drop-off zone in Cluster Transition or Residential-business in the South-West, they tend to be non-transit-competing. However, having a drop-off zone in Business-residential or Rural in the South-East community slightly increases the probability of being transit-competing.

The demand-based communities of the pick-up zone and drop-off zone interact with each other (Fig. 9G), the transit-competing trips are slightly more likely to happen within each community, likely due to better connections by PT, while if a trip is from the South-East to the North, it is more likely to be non-transit-competing.

Besides the demand-based community, the functional clusters of the pick-up and drop-off zones also interact with each other (Fig. 9H). The trips from Cluster Business-residential or Rural to Cluster Outer-residential, Business-residential, or Rural tend to be transit-competing. By adding how the trip direction impacts the effect of the built

environment on whether a trip is transit-competing, this interaction term complements the insights from the single-feature effect of the functional cluster (Fig. 8E and F).

4.3. Cases of transit-competing trips

Based on the model output (Fig. 8–9), we find some typical situations in which the probability of ride-sourcing being transit-competing is above average. For the single-feature components, the competition tends to happen when TT by ride-sourcing <15 min, TT ratio $\in [1.8, 3.5]$, # of boardings $\in$ {3, 4, 5, 6}, Weather $\in$ {Fog, Mist, Rain}, Functional cluster (pick-up & drop-off zone) is Cluster Centre, and transit-stop density $\in [13, 50]$.

For the feature interaction components, here is a brief summary: (1) TT ratio >2 and # of boardings = 2; (2) Weather is clear and TT ratio >3.4; (3) # of boardings $\in$ {4, 5, 6} and Functional cluster of drop-off zone in Cluster Transition or Outer-residential; (4) Transit-stop density >20 and Functional cluster of drop-off zone in Cluster Business-residential; (5) Demand-based community of pick-up and drop-off zones in South-East; (6) Functional cluster of pick-up and drop-off zones in Cluster Business-residential or Rural to Cluster Outer-residential, Business-residential, or Rural.

Fig. 10 shows two cases: (A) TT ratio >2 and # of boardings = 2; (B) Transit-stop density >20 and Functional cluster of drop-off zone in Cluster Business-residential, regarding the top five pairs of pick-up and drop-off zones that meet the case conditions. For the trips for which the travel time by PT excl. access/egress walking is more than two times as long as TT by ride-sourcing and a transfer is needed, the presented top origin-destination pairs are concentrated in the central and eastern areas (Fig. 10A). Despite the short distances between them, the need of transferring once diminishes the competitiveness of the PT alternative. These trips imply an unmet PT demand presumably due to multiple transfers despite relatively easy access to the nearby PT stations. Moreover, the short travel time by ride-sourcing can be attributed to the easy access to the major highway entrance and exit. However, these trips still require relatively long travel time by PT, despite the short walking distance (making them transit-competing). The ride-sourcing trips in Fig. 10B mostly have a Business-residential drop-off zone located in the South-Eastern area and its transit-stop density is fairly high indicating easy access to PT. In this drop-off zone, there locates the Chengdu East railway station.

5. Discussion

In this study, a ride-sourcing trip is considered transit-competing if its PT alternative requires less than 800 m of access and egress walking and its departure time is during 6 am–11 pm. By comparing and classifying features of transit-competing and non-transit-competing ride-sourcing trips, this study seeks to identify factors that could alleviate the competition between ride-sourcing and PT and make them more complementary. Whether a ride-sourcing trip is transit-competing is trained and predicted using the trip attributes and the characteristics of the built environment. The results demonstrate the effectiveness of the enhanced GAM: The impact of single features and the interactions between them are clearly revealed and visualised.

Of the 4.27 million ride-sourcing trips, 48.2% compete with PT according to the definition above. Despite the binary simplification of the relationship, this number suggests that a considerable share of ride-sourcing trips can potentially be done by taking PT. This is consistent with previous studies (e.g., Wang and Ross, 2019; Barajas and Brown, 2021). In this study, the distribution of pick-up and drop-off hot spots for non-transit-competing trips has a greater spread than the ones for transit-competing trips, presumably because some of the grid cells do not have good access to the nearby PT stations. It is worth noting that in this study, the definition of transit-competing only takes walking distance and departure time into consideration, while socio-demographic dimensions and trip purposes have also been found to be important.

Despite the short walking distance, there are other circumstances where PT is not an alternative mode for some travellers. For instance, both categories of ride-sourcing trips have drop-off hot spots at the international airport and the railway station (Fig. 5). This confirms a previous study in the US (Zhen, 2015) where 40% of the survey respondents stated that they mainly use ride-sourcing to get to or from the airport. The case shown in Fig. 10B is consistent with the results in Fig. 5, indicating that the willingness to take PT for long-distance trips is less affected by transit access. In other words, despite good PT access to the airport and railway station for those transit-competing trips, some travellers may be unwilling to take PT when, for instance, carrying

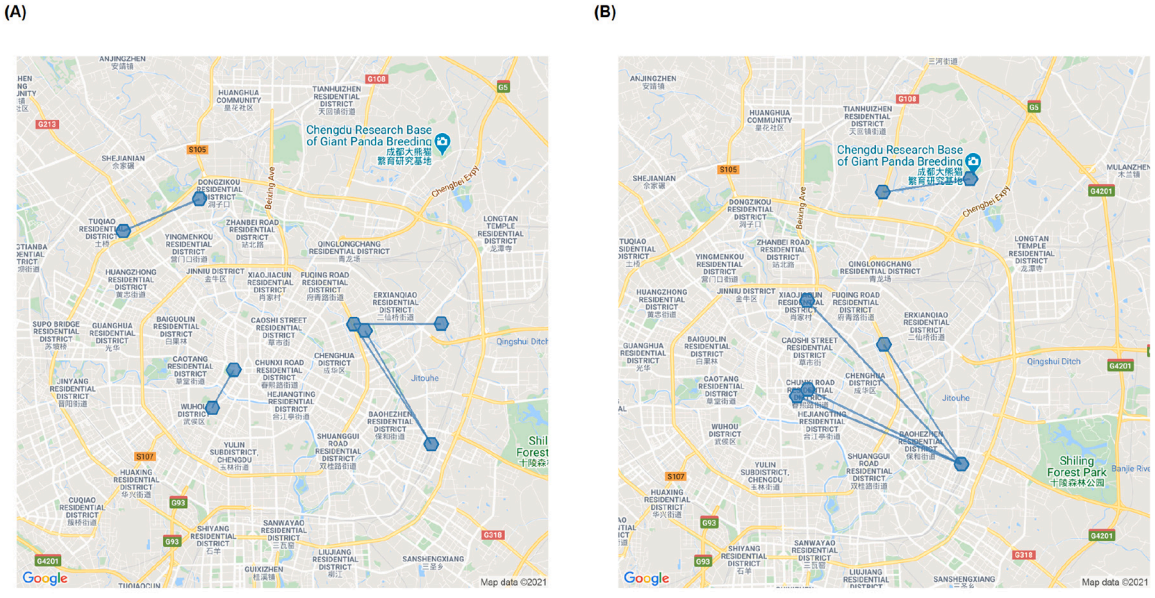


Fig. 10. Top pick-up and drop-off zone pairs of transit-competing trips for two cases. (A) TT ratio >2 and # of boardings = 2 (637 trips encompassed). (B) Transit-stop density >20 and Functional cluster of drop-off zone in Cluster Business-residential (1436 trips encompassed).

luggage. One explanation for this is the consideration of vehicle comfort (Redman et al., 2013); with a long journey by air or rail ahead, the passengers value the access trip more than usual.

The explored trip attributes **interactively** impose a significant impact on whether a ride-sourcing trip is transit-competing. **Travel time** is one of the most critical factors in the choice of travel mode. This study finds that short trips (<15 min by ride-sourcing) tend to be transit-competing (Fig. 8A). This can be explained by walking time being perceived negatively especially for short journeys (Walle and Steenberghen, 2006). For these transit-competing short trips, walking would take up a big share of total travel time were the trip done by PT. If one wants to ease the competition between ride-sourcing and PT for a better mix of modes, this observation suggests decreasing the travel time by PT, especially for those ride-sourcing trips for which TT by ride-sourcing is less than 15 min.

Taking out the factor of walking, **the TT ratio** is an indicator used to reflect the in-vehicle time disparity between PT and ride-sourcing (Fig. 8B). If the TT ratio is above 1.5, a ride-sourcing trip is more likely to be transit-competing. Given people's preferences for travelling faster where the PT alternative was on the slow side (Redman et al., 2013), the ride-sourcing service in the study area to some degree fills the demand that requires too long travel time relative to ride-sourcing.

Besides walking and travel time, **transfers** (Fig. 8C) are also perceived negatively. A study indicates that a transfer can be equivalent to 5–20 in-vehicle time minutes (Walle and Steenberghen, 2006). The more transfers are needed, the more likely a trip is transit-competing. This suggests that ride-sourcing covers the travel demand where the transfers are too many despite short access and egress walking distances. What had not been observed is that the in-vehicle time disparity between ride-sourcing and PT interacts with the number of transfers (Fig. 9A): when at least one transfer is needed, the competition tends to happen even for those trips of little disparity. This highlights the penalty of transferring which makes PT less competitive than ride-sourcing. Therefore, the strategy to ease the competition can be decreasing the number of transfers.

The influence of **weather** on the ridership of public transit has been studied extensively (Zhou et al., 2017). Here, we reveal the effect that weather has on the relationship between ride-sourcing and its PT alternative. Poor weather conditions such as fog, mist, and rain tend to ease the competition between ride-sourcing and PT. This is consistent with previous findings that rainfall would typically increase the use of public transit as walking or driving might be quite difficult under such conditions (Zhou et al., 2017). It means that the willingness of taking PT increases under these weather conditions resulting in less transit-competing ride-sourcing trips. On the other hand, as suggested by the effect of the interaction between weather and travel time ratio, if the weather is clear, the competition tends to happen if the TT ratio is greater than 2.1 (Fig. 9C). For the poor weather conditions, sensitivity to travel time disparity is low.

The **built environment**, as characterised by the identified functional clusters and transit-stop density, affects whether a ride-sourcing trip is transit-competing (Fig. 7). Higher diversity and density of land use encourage the choice of non-driving modes (Zhang, 2004). However, despite the high diversity and density of land-use patterns in Cluster Centre (better access to PT as well), the ride-sourcing trips there have a slightly higher tendency to compete with PT (Fig. 8E and F). On the flip side, the ride-sourcing trips in the other areas accounting for around 50%, such as Outer-residential and Rural where the land-use density/-diversity is not as great as the central city, are less transit-competing. A study has found that if the ride-sourcing trips substitute for PT in lower-density areas, which is observed in the study area, the energy efficiency gains seem to be higher (Becker et al., 2020). This implies that the role of ride-sourcing in Chengdu is leaning towards the complementary side to the PT system. Consistent with the higher probability of competing with the transit in the centre city where the transit access is good, we observe a negative impact of **transit-stop density** on the competition (Fig. 8G and H); the better the transit access, the more likely a ride-sourcing trip is transit-competing. A similar relationship has also been found by Barajas and Brown (2021), who find ride-sourcing services are not filling the demand gap in the areas of low transit-stop density.

When affecting the probability of ride-sourcing being transit-competing, the factor of the built environment interacts with some trip attributes. Clusters Outer-residential and Transition are in the middle area between Chengdu city centre and the surrounding cities' centres, and they have a higher probability of generating transit-competing trips only if the PT alternative requires multiple transfers (Fig. 9D). These zones have a low density of economic activity according to their lower number of various POIs compared with the rest of the study area (Fig. 6A). This suggests that there is room for improvement of PT in Clusters Outer-residential and Transition by reducing the transfer inconvenience by increasing connectivity between the central city and these areas.

Regarding the built environment, the trips attracted to the zones of moderate to high transit-stop density in Cluster Business-residential are more likely to be transit-competing (Fig. 9E). One explanation could be that given that Cluster Business-residential features a large number of business and much real estate, this tendency is due to a higher probability of business trips instead of private ones. As suggested by a previous survey study (Alemi et al., 2018), the respondents who report higher numbers of long-distance business trips are also more likely to have used ride-sourcing services.

In addition to the model results, the two cases present a way of extracting regional insights on the situations with an increased probability of competition. For a more sustainable mix of transport modes, the analysis contributes to planning by directly looking at ride-sourcing demand while considering PT as an alternative.

We identify a few key points for transport planning and policymaking based on the findings. For making PT more competitive, better PT services that provide access to the international airport are needed, given the airport being the hotspot of ride-sourcing trips that oftentimes require lengthy walking by PT. Moreover, ride-sourcing tends to compete with PT for short trips below 15 min, especially those of great travel time gap between the two modes; PT planners should look into where and when these ride-sourcing trips distribute to guide the future expansion of PT networks to optimise the PT coverage. Such planning work could also consider increasing the connectivity between the functional urban regions, Outer-residential and Transition, and the rest of the study area. Of course, PT services cannot and should not cover every corner of cities. For policymaking to reduce the GHG emissions from transport systems, employers and ride-sourcing platforms could incentivise the ride-sourcing trips that fill the gaps in the PT services, e.g., the trips that take a long time for PT or require lengthy walking and transfers connecting to suburban areas. At last, one could better combine the travel information of ride-sourcing and PT to increase the convenience of using these two modes jointly for the first- and last-mile situations.

5.1. *Limitations and future work*

The study has two main limitations. The first limitation is about the definition of transit-competing. The daytime (6 am–11 pm) ride-sourcing trips that have a PT alternative with a walking distance of less than 800 m for each of access and egress are defined as transit-competing, and all others are non-transit-competing, although the walking distances and time of day are not the only indicators affecting PT adoption in reality. Further explorations could for instance examine a varying walking-distance threshold. Furthermore, this distance-based definition does not include other constraints, such as trip purpose and socio-demographics of riders, known to be important to mode choice. The open dataset used for the analysis was created passively, so it is not possible to access that information. However, this study presents a

replicable framework that utilises open sources to enrich such a dataset and applies cross-disciplinary tools that contribute to intelligible outcomes. Open, large, but incomplete data will be more and more available. Such data benefit from the framework proposed in this study.

The second limitation pertains to the discussion of the relationship between ride-sourcing and PT. Though this study aims to discuss ride-sourcing compared with PT, it is centred on ride-sourcing due to the lack of concurrent PT trips. This makes this study explore the relationship between ride-sourcing and PT in a virtual space that supposes the following: What if these ride-sourcing trips were done by taking PT? Due to the lack of PT trips, this study can only focus on the absolute number of ride-sourcing trips instead of its relative percentage among all trips. To better inform policymaking, PT big trip data need to be collected from other sources, for instance, smart cards.

6. Conclusions

This study explores the competition between ride-sourcing and PT through the lens of big data analysis. The contributions pertain to methodological and empirical aspects. Methodologically, we apply a data fusion framework without involving empirical PT trip records. Applying a glass-box model on the enriched ride-sourcing trip data provides a good overview of not only the main factors affecting the relationship between ride-sourcing and PT, but also the interactions between those factors; the latter is lacking in the literature. From the perspective of gaining new knowledge, data from developing countries are generally under-exploited to discuss the relationship between ride-sourcing and PT. The obtained insights of this study are useful to guide the local transport planning and they also contribute to an improved big picture of how global cities are experiencing ride-sourcing.

Spatio-temporally, the travel demand for transit-competing trips largely overlaps with that for non-transit-competing trips. The transit-competing trips account for 48.2% of the total trip records studied. Competition is more likely to happen when the travel time by ride-sourcing <15 min or the travel time by PT is disproportionately longer than ride-sourcing (in-vehicle travel time ratio >1.8). Requiring multiple transfers is also associated with the competition between ride-sourcing and PT, especially for the trips within the transition area between the central city and the outskirts. Poor weather conditions, such as rain, tend to ease the competition between ride-sourcing and PT, where the ride-sourcing users seem to be less sensitive to the travel time disparity between the two modes. Functional cluster of urban regions is the most important factor in determining the relationship between the two modes. Both low density and low diversity of land use are associated with a lower probability of generating transit-competing trips. The better the transit access, the more likely a ride-sourcing trip is transit-competing, especially for the areas featuring a large number of companies and real estate.

Some recommendations for transport planning based on the main findings are to: (1) Improve PT services that provide access to the international airport; (2) Expand PT networks guided by the transit-competing ride-sourcing trips featuring short travel time but a big travel time disparity between the two modes; (3) Increase the connectivity between the functional urban regions, Outer-residential and Transition, and the rest of the study area; (4) Incentivise the ride-sourcing trips that fill the gaps in the PT services where PT takes a long time or requires lengthy walking and transfers connecting to sub-urban areas; (5) Better combine the travel information of ride-sourcing and PT for travellers for the first- and last-mile issues.

Funding

The author acknowledges the financial support of the Swedish Research Council for Sustainable Development (Formas, project number 2016-01326).

Conflicts of interest

The author declares no competing financial or non-financial interests.

Availability of data and material

The aggregate data that support the findings of this study are available upon request from the corresponding author. The original data set can be requested online via DiDi Chuxing GAIA Open Dataset Initiative.

Authors' contributions

Yuan Liao designed the study, analysed the data, and wrote the manuscript.

Acknowledgements

This research is funded by the Swedish Research Council Formas (Project Number 2016-1326).

References

Aarhaug, J., Olsen, S., 2018. Implications of ride-sourcing and self-driving vehicles on the need for regulation in unscheduled passenger transport. Res. Transp. Econ. 69, 573–582.
Alemi, F., Circella, G., Handy, S., Mokhtarian, P., 2018. What influences travelers to use uber? Exploring the factors affecting the adoption of on-demand ride services in California. Travel Behav. Soc. 13, 88–104.
Allen, M., Cervo, D., 2015. Multi-Domain Master Data Management: Advanced MDM and Data Governance in Practice. Morgan Kaufmann.
Baidu Maps, 2020a. Place API. http://lbsyun.baidu.com/index.php?title=webapi/guide/webservice-placeapi (May).
Baidu Maps, 2020b. Transit API. https://lbsyun.baidu.com/index.php?title=webapi/direction-api-v2 (November).
Barabási, A.-L., et al., 2016. Network Science. Cambridge University Press.
Barajas, J.M., Brown, A., 2021. Not minding the gap: does ride-hailing serve transit deserts? J. Transp. Geogr. 90 (January (102918)), 1–14.
Becker, H., Balac, M., Ciari, F., Axhausen, K.W., 2020. Assessing the welfare impacts of shared mobility and mobility as a service (maas). Transp. Res. Part A Policy Pract. 131, 228–243.
Brustein, J., 2016. Uber and lyft Want to Replace Public Buses. https://www.bloomberg.com/news/articles/2016-08-15/uber-and-lyft-want-to-replace-public-buses. , Bloomberg.
Burdziej, J., 2019. Using hexagonal grids and network analysis for spatial accessibility assessment in urban environments-a case study of public amenities in Toruń. Misc. Geogr. 23 (2), 99–110.
California Air Resources Board, 2019. Sb 1014 Clean Miles Standard 2018 Base-Year Emissions Inventory Report. California Air Resources Board, United States. Tech. Rep.
Cervero, R., Kockelman, K., 1997. Travel demand and the 3ds: density, diversity, and design. Transp. Res. Part D Transp. Environ. 2 (3), 199–219.
Chengdu Bureau of Statistics, 2020. 2019 Statistical Bulletin of Chengdu's National Economic and Social Development. http://www.cdstats.chengdu.gov.cn (March).
City of New York, 2020. TLC Trip Record Data. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page (May).
Cordera, R., Coppola, P., dell'Olio, L., Ibeas, A., 2017. Is accessibility relevant in trip generation? Modelling the interaction between trip generation and accessibility taking into account spatial effects. Transportation 44 (6), 1577–1603.
Currie, G., 2018. Lies, damned lies, avs, shared mobility, and urban transit futures. J. Public Transp. 21 (1), 19–30.
De Vos, J., Mokhtarian, P.L., Schwanen, T., Van Acker, V., Witlox, F., 2016. Travel mode choice and travel satisfaction: bridging the gap between decision utility and experienced utility. Transportation 43 (5), 771–796.
DiDi Chuxing, 2020a. Didi Chuxing – About Us. https://www.didiglobal.com/about-didi/about-us (May).
DiDi Chuxing, 2020b. Didi Chuxing Gaia Open Dataset Initiative. https://gaia.didichuxing.com (May).
Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitao, P.J., et al., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36 (1), 27–46.
Ewing, R., Cervero, R., 2010. Travel and the built environment: a meta-analysis. J. Am. Plan. Assoc. 76 (3), 265–294.
Gao, S., Janowicz, K., Couclelis, H., 2017. Extracting urban functional regions from points of interest and human activities on location-based social networks. Trans. GIS 21 (3), 446–467.

Getis, A., Ord, J.K., 2010. The analysis of spatial association by use of distance statistics. Perspectives on Spatial Data Analysis. Springer, pp. 127–145.
Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models, vol. 43. CRC Press.
Hochmair, H.H., 2016. Spatiotemporal pattern analysis of taxi trips in New York city. Transp. Res. Record 2542 (1), 45–56.
Hu, S., He, Z., Wu, L., Yin, L., Xu, Y., Cui, H., 2020. A framework for extracting urban functional regions based on multiprototype word embeddings using points-of-interest data. Comput. Environ. Urban Syst. 80, 1–15.
Huang, L., Yang, Y., Gao, H., Zhao, X., Du, Z., 2018. Comparing community detection algorithms in transport networks via points of interest. IEEE Access 6, 29729–29738.
IEA, 2012. CO2 Emissions From Fuel Combustion. IEA, Paris. http://data.iea.org (Tech. Rep.).
Jain, A.K., 2010. Data clustering: 50 years beyond k-means. Pattern Recognit. Lett. 31 (8), 651–666.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning, vol. 112. Springer.
Jiang, Y., Zegras, P.C., Mehndiratta, S., 2012. Walk the line: station context, corridor type and bus rapid transit walk access in Jinan, China. J. Transp. Geogr. 20 (1), 1–14.
Kamga, C., Yazici, M.A., Singhal, A., 2015. Analysis of taxi demand and supply in New York city: implications of recent taxi regulations. Transp. Plan. Technol. 38 (6), 601–625.
Liu, X., Gong, L., Gong, Y., Liu, Y., 2015. Revealing travel patterns and city structure with taxi trip data. J. Transp. Geogr. 43, 78–90.
Lou, Y., Caruana, R., Gehrke, J., Hooker, G., 2013. Accurate intelligible models with pairwise interactions. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 623–631.
Mason, R.L., Gunst, R.F., Hess, J.L., 2003. Statistical Design and Analysis of Experiments: With Applications to Engineering and Science, vol. 474. John Wiley & Sons.
McCullagh, P., 2018. Generalized Linear Models. Routledge.
Molnar, C., 2020. Interpretable Machine Learning. Lulu.com.
Murphy, J.J., Allen, P.G., Stevens, T.H., Weatherhead, D., 2005. A meta-analysis of hypothetical bias in stated preference valuation. Environ. Resour. Econ. 30 (3), 313–325.
Narayan, J., Cats, O., van Oort, N., Hoogendoorn, S., 2019. Does ride-sourcing absorb the demand for car and public transport in amsterdam?. In: 2019 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS). IEEE, pp. 1–7.
National Bureau of Statistics of China, 2020. China Statistical Yearbook 2020. http://www.stats.gov.cn/tjsj/ndsj/2020/html/E1620.jpg (January).
Nori, H., Jenkins, S., Koch, P., Caruana, R., 2019. Interpretml: A Unified Framework for Machine Learning Interpretability. arXiv:1909.09223 (arXiv preprint).
Qian, X., Ukkusuri, S.V., 2015. Spatial variation of the urban taxi ridership using gps data. Appl. Geogr. 59, 31–42.
Rayle, L., Dai, D., Chan, N., Cervero, R., Shaheen, S., 2016. Just a better taxi? A survey-based comparison of taxis, transit, and ridesourcing services in San Francisco. Transp. Policy 45, 168–178.
Reck, D.J., Axhausen, K.W., 2019. Ridesourcing for the first/last mile: how do transfer penalties impact travel time savings?. In: International Scientific Conference on Mobility and Transport. (mobil. TUM 2019).
Redman, L., Friman, M., Gärling, T., Hartig, T., 2013. Quality attributes of public transport that attract car users: a research review. Transp. Policy 25, 119–127.
Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.
Ryan, S., Frank, L.F., 2009. Pedestrian environments and transit ridership. J. Public Transp. 12 (1), 3.
Sarker, R.I., Mailer, M., Sikder, S.K., 2019. Walking to a public transport station: empirical evidence on willingness and acceptance in Munich, Germany. Smart Sustain. Built Environ.
Schafer, A., Victor, D.G., 2000. The future mobility of the world population. Transp. Res. Part A Policy Pract. 34 (3), 171–205.
Schäfer, A.W., Yeh, S., 2020. A holistic analysis of passenger travel energy and greenhouse gas intensities. Nat. Sustain. 1–4.
Shaheen, S., Cohen, A., Zohdy, I., et al., 2016. Shared Mobility: Current Practices and Guiding Principles. Federal Highway Administration, United States. Tech. Rep.
Shared-use Mobility Center, 2020. What is Shared Mobility. https://sharedusemobilitycenter.org/what-is-shared-mobility/ (May).
Shirgaokar, M., 2018. Expanding seniors' mobility through phone apps: potential responses from the private and public sectors. J. Plan. Educ. Res. 1–11.
Sobolevsky, S., Campari, R., Belyi, A., Ratti, C., 2014. General optimization technique for high-quality community detection in complex networks. Phys. Rev. E 90 (012811), 1–8.
Tolley, R., 2016. Supporting walking in cities-best practice around the world. In: Walk the City International Conference, vol. 26. Stavanger, October.
Walle, S.V., Steenberghen, T., 2006. Space and time related determinants of public transport use in trip chains. Transp. Res. Part A Policy Pract. 40 (2), 151–162.
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 (3), 1051–1072.
Welch, T.F., Gehrke, S.R., Widita, A., 2020. Shared-use mobility competition: a trip-level analysis of taxi, bikeshare, and transit mode choice in Washington, DC. Transp. A Transp. Sci. 16 (1), 43–55.
Winter, K., Cats, O., Correia, G., van Arem, B., 2018. Performance analysis and fleet requirements of automated demand-responsive transport systems as an urban public transport service. Int. J. Transp. Sci. Technol. 7 (2), 151–167.
Yan, X., Levine, J., Zhao, X., 2019. Integrating ridesourcing services with public transit: an evaluation of traveler responses combining revealed and stated preference data. Transp. Res. Part C Emerg. Technol. 105, 683–696.
Yu, H., Peng, Z.-R., 2019. Exploring the spatial variation of ridesourcing demand and its relationship to built environment and socioeconomic factors with the geographically weighted poisson regression. J. Transp. Geogr. 75, 147–163.
Zhang, M., 2004. The role of land use in travel mode choice: evidence from Boston and Hong Kong. J. Am. Plan. Assoc. 70 (3), 344–360.
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. Transp. Geogr. 70, 78–90.
Zhen, C., 2015. Impact of Ride-Sourcing Services on Travel Habits and Transportation Planning. University of Pittsburgh. Ph.D. Thesis.
Zhou, M., Wang, D., Li, Q., Yue, Y., Tu, W., Cao, R., 2017. Impacts of weather on public transport ridership: results from mining data from different sources. Transp. Res. Part C Emerg. Technol. 75, 17–29.