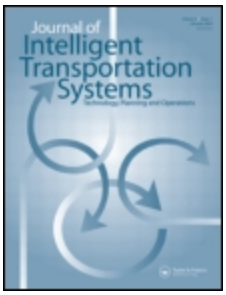

Identifying critical transfer zones to coordinate transit with on-demand services using crowdsourced trajectory data

Journal:	*Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*
Manuscript ID	GITS-2022-0210.R2
Manuscript Type:	Research Article
Keywords:	trajectory data, transit, on-demand services, first/last mile, Ridesharing

Identifying critical transfer zones to coordinate transit with on-demand services using crowdsourced trajectory data

Abstract

This study develops a data-driven approach for identifying critical transfer zones in the city to facilitate the coordination of transit and emerging on-demand services. First, the methods convert the trajectories into a 3D grid with an optimal cube size. Built upon that, we zoom in and study the trajectory density of each mode in a cube and present the results by heatmaps. After that, we zoom out and aggregate those cube information fragments through the clustering algorithms to explore two critical patterns: the ridesharing swarm (RS) zones where many ridesharing trips go through, and the "sandwich pattern" zones where a transit trajectory dominant zone is sandwiched by two ridesharing trajectory dominant zones. Our numerical analysis confirms that these RS zones are well correlated to the promising areas/corridors for integrating transit and on-demand services; the "sandwich patterns" help discover first/last mile (FLM) zones. Last, we further develop a two-channel deep learning network to predict the variation of the FLM gaps so that adaptive services can be planned. A case study based on the field data of the second ring region of Chengdu, China confirms the effectiveness and capability of our analysis approach.

Keywords: trajectory data, ridesharing, transit, on-demand services, first/last mile.

Introduction

The benefits of coordinating transit systems and emerging on-demand services, such as microtransit, micro-mobility, ridesharing services, and ride-hailing, have been recently recognized (Boarnet et al., 2017; Koffman, 2004). The potential schemes seek to integrate flexibility into the current transit service to better meet dynamic travel demands. To date, even though much research (Aldaihani et al., 2004; Fu, 2002; J.-Q. Li et al., 2012; X. Li & Quadrifoglio, 2010; Quadrifoglio & Li, 2009) has focused on how to manage and operate such hybrid mobility services, little attention has been paid to identifying critical transfer locations and times for the better coordination (Velaga et al., 2012). Specifically, it is about discovering the most appropriate zones or locations in a city (a.k.a., hotspots, critical corridors/connections (see Figure 1)) where on-demand services can best supplement to given transit services. Clearly, it is not trivial to narrow down all those critical locations in a big city. This unique niche involves two types of mobility services with different service flexibility and demand variations. Therefore, it calls for new research co-considering the mobility patterns associated with both modes.

Most existing studies only investigated the service and predicted passenger demand of a single mode (either transit or on-demand services). For example, using the transit ridership data, demographic survey, and land use data, many studies (Boyle, 2006; Hashemian, 2002; Huang, 1996; Nazem et al., 2011; Roberts, 1985; Sung et al., 2014; Trépanier et al., 2007) estimated transit demands only at some candidate locations for planning transit routes, and these estimations mainly reflected long-term trends but not short-term changes like weekly or monthly. Even though it has been recommended that transit network redesign should consider how the emerging modes can complement the transit services (Johnson et al., 2020), we still lack efficient methods to identify the candidate zones and then develop efficient operations. On the other hand, extensive ridesharing[1] trajectory data have been well collected and used as valuable data to predict the dynamic ad hoc demands (Faghih et al., 2019; Liu et al., 2019; Xu et al., 2017; K. Zhang et al., 2019), mainly for improving ridesharing services, i.e., Uber and taxi. Overall, the state-of-the-art literature indicates that the transit and ridesharing service data are often individually analyzed with separate objectives for the respective modes. Thus, existing studies have not provided a comprehensive understanding of the critical zones that are suitable for integrating public transit and on-demand services.

Given the above issues, this study proposes a data-driven approach using large-scale crowdsourced transit and ridesharing trajectory data to identify the spatiotemporal service gaps (also called critical transfer

[1] The existing study (https://www.ecolane.com/blog/ride-hailing-vs.-ride-sharing-the-key-difference-and-why-it-matters) shows that TNC such as Uber and Lyft provide ridesharing services, but majority of the trips are ride-hailing services. Therefore, this study uses ridesharing to broadly cover carpool and ride-hailing services. Later, you will see that our study focuses on analyzing the trajectory of the ridesharing trips from the supply side. A carpool service will be separated into multiple trips, each with different origin and destination, while a ride-hailing corresponds to one trajectory. Therefore, we do not differentiate this terminology (ridesharing or ride-hailing) hereafter in this study.

zones in this paper) for implementing new transfer mobility services to promote the cooperation between public transit and on-demand services. To do that, this study thinks that the competition and/or complement between transit and ridesharing services indicate a strong correlation between their services reflected by their routes. Therefore, this trajectory data generated from a large scale of transit and ridesharing services represent the crowdsourced data, which can be used to detect those critical transfer zones for their cooperation. For example, by jointly analyzing many pick-up and drop-off locations of ridesharing services and the distribution of transit stops and schedules, we can find some critical zones where a high volume of ridesharing orders is present but with low transit services. Then, these zones are the critical candidate hotspots for setting up on-demand services such as microtransit (Xu et al., 2017) for strengthening the connection to the nearby transit network. More importantly, by uniting and then analyzing large-scale trajectories generated by the services of both ridesharing and transit vehicles (e.g., bus and metro), we expect to find some critical corridor zones involving multiple roads, which are densely passed by ridesharing vehicles but not transit vehicles. Then, we can conjecture that those corridors are good candidate zones to set up new transit connections/links in transit network redesign. In short, we think that uniting and then analyzing large-scale ridesharing and transit trajectory data will help discover those critical zones (see Figure 1), which are promising candidates to be considered by planners for cooperating transit and on-demand services.

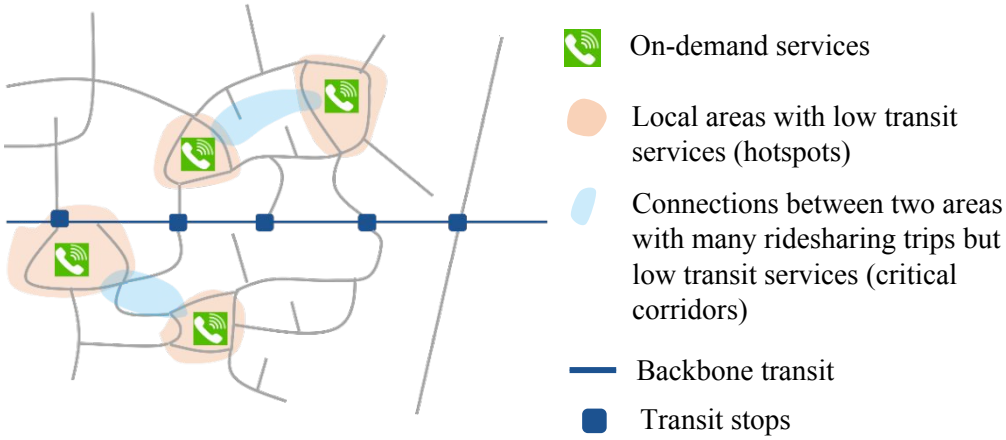

Figure 1. Hotspots and critical corridors.

However, the proposed data analysis is not trivial. It raises new research difficulties. For example, the transit trips and ridesharing trajectories are collected with different spatiotemporal resolutions. When we put them in a spatiotemporal (3D) space, those data are non-additive and deter many quantitative approaches to start directly. More importantly, even though the pick-up and drop-off data are relatively easy to be aggregated to passenger demand, it is not apparent to infer critical transfer/corridor zones from trajectory curves, given existing literature does not accurately define their features yet. We will develop novel approaches to address these issues.

Overall, this study develops a new data analysis approach to analyze the large-scale crowdsourced trajectory data collected from ridesharing and transit services, aiming to find the critical transfer zones for cooperating emerging on-demand services with transit services. Specifically, we contribute the main methods as follows. (i) This study develops a new data presentation approach, which meshes the trajectory curves within an optimal 3D grid. The size of the cubes is determined by the optimization models which balance the data resolution and computation load in further data analysis. (ii) This study develops an innovative approach to analyze entire trip data beyond their pick-up and drop-off information. Specifically, we employ pattern recognition to discover the critical zones shown in the heatmaps (i.e., trajectory intensity map) generated from the 3D representation, and then learn the FLM (first-and-last) mile zones evolution. (iii) In particular, we investigate two interesting and unique patterns reflected by the large-scale trajectory data: RS zones and "sandwich" patterns. They both present great values to infer the critical transfer zones for integrating transit and on-demand services. (iv) We analyze a set of field trajectory data collected from Chengdu city in China and validated the effectiveness of our approaches. Note that the results found by our approaches can be further sharpened by integrating land-use data, demographic data, micro-mobility data, etc. To the best of our knowledge, we are one of the first studies working on united trajectory data collected from different modes. This is also a pioneering study investigating the critical transfer zones for the cooperation of public transit and on-demand services. More importantly, the crowdsourced data analysis approach can be extended to study other trajectory data, such as bike-sharing and private vehicle trajectory

data, to provide a thorough understanding of mobility services over a city network. These contributions together benefit the development and operations of hybrid mobility services in urban areas.

The efforts of this study are organized by the following structure. The next section reviews the most relevant literature and highlights the unique contribution of this study. Following that, we formally define the problem, and develop our methods to analyze the crowdsourced trajectory data. We further conduct the case study to validate the effectiveness of our approach, and then summarize the entire study in the conclusion section.

Literature review

Given the scope of our study, we will review the literature about data-driven transit and ridesharing demand prediction and service design. We first review the relevant studies on transit demand prediction. Boyle shows that few studies use large-scale trajectory data as the crowdsourced data from its competitive modes to infer transit service needs (Boyle, 2006), even though various data has been analyzed, including transit ridership (Fang et al., 2018; Noursalehi et al., 2018), demographic (Nazem et al., 2011; Roberts, 1985), land use (Hashemian, 2002; H. Huang, 1996; Jun et al., 2015; Sung et al., 2014) and O-D survey data (Chatterjee & Venigalla, 2004). For example, Nazem et al., (Nazem et al., 2011) analyzed the travel patterns of different demographic classes to understand the relationship between transit ridership and demographics. Sung et al., employed spatial regression analysis to investigate the impact of land use on the rail transit ridership in the city of Seoul (Sung et al., 2014). Jun et al., applied a multinomial logit model to analyze how land use and demographic characteristics affect transit ridership (Jun et al., 2015). In recent years, the ridership data collected by the Automated Fare Collection (AFC) system has been used to capture the variation of the transit demand, especially for railway systems (Fang et al., 2018). For example, based on AFC data, Noursalehi et al., developed the state-space model to predict the real-time subway demand, considering the impact of special events (Noursalehi et al., 2018). Even though those data are directly oriented towards passengers' features and mobility needs, and can help predict transit demand at some locations well, they provide limited insights for the integration of transit with other mobility services.

The review recognizes that extensive studies have analyzed the ridesharing data with various purposes, such as predicting ridesharing demand, providing optimal routes (Yuan et al., 2010), predicting traffic conditions (Castro et al., 2012), rebuilding routable city road map (Cao & Krumm, 2009), or detecting urban planning flaws (Zheng et al., 2011). We briefly discuss several of them working on predicting ridesharing service needs. Based on Uber pick-up data, Faghih et al., applied the Least Absolute Shrinkage and Selection Operator (LASSO) spatial-temporal autoregressive model to predict the Uber demand (Faghih et al., 2019). Xu et al., fed the taxi pick-up and drop-off data in New York City into a long short-term memory (LSTM) neural network to forecast the future taxi requests (Xu et al., 2017). Zhou et al., employed the convolutional LSTM (ConvLSTM) to capture the spatiotemporal relationship of taxi and bike-sharing demand data in New York for a short-term demand prediction (Zhou et al., 2018). Zhang et al., developed an end-to-end multi-task learning temporal convolutional neural network to predict the short-term ridesharing demand and compared its performance with the state-of-the-art deep learning approaches (K. Zhang et al., 2019). It should be noted that none of the above studies focus on inferring potential critical transfer zones suitable to connect transit routes with on-demand services. This study seeks to fill in this gap.

We also notice that many studies develop data-driven approaches to estimate mobility demand for the transit service design by using the data outside of transit systems such as taxi GPS data, mobile phone GPS data and bikesharing GPS data. For example, Chen et al., analyzed taxi GPS traces to uncover the areas with dense pick-up/drop-off orders as candidate bus stops (Chen et al., 2013). Wang et al., designed a taxi-sharing and subways (TSS) system. By solving a matching model upon the taxi GPS data, the route plan of TSS is determined to maximize the number of participants (Wang et al., 2021). Bastani et al., applied an agglomerative clustering algorithm to cluster taxi trips based on their origin/destination and then developed a routing algorithm to identify optimal routes for a flexible mini-shuttle that connects multiple taxi trip clusters (Bastani et al., 2011). Berlingerio et al., developed a system, AllAboard, to extract origin-destination flows and sequential travel patterns from mobile data (Berlingerio et al., 2013). Then, new routes were added to an existing transit network to accommodate the identified O-D flows. Hadjidimitriou et al., used mobile data to estimate the potential time-variant O-D flows and then compared it with the field transit service so that they could identify the unmatched demands and propose new transit routes (Hadjidimitriou et al., 2020). Shu et al., proposed a data-driven method and used bikesharing data to design shuttle services to improve the efficiency of last mile transportation (Shu et al., 2021). The above review shows that most of these data-driven approaches analyzed the pick-up and drop-off data rather than the entire trajectories. Therefore, the collective information involved in the trips of the ridesharing services is not well investigated. Given trajectory data are non-additive and more complicated as compared to the pick-

up and drop-off data, this study contributes new analysis approaches to address this difficulty.

From the application view, this study can narrow down the searching space for those candidate critical zones, which should be considered in the transit redesign for integrating transit systems with emerging on-demand services. The concept of the hybrid system has been proposed for over decades and many operation models have been proposed to show its merit (Berrada & Poulhès, 2021; Maheo et al., 2019; Mounce et al., 2018; Rahimi & Dessouky, 2001; Stiglic et al., 2018; Teal, 1994). For example, Luo and Nie found that the hybrid system that mixes ride-pooling and fixed-route services can improve the overall system efficiency while maintaining the economy of scale in transit design (Luo & Nie, 2019). Grahn et al, simulated daily operations for an existing first-mile last-mile mobility service and indicated that added flexibility of the hybrid service (using shuttles and TNCs) improved service performance by 7.7% (Grahn et al., 2022). However, existing studies show that only a small percentage of transit agencies (Potts et al., 2010) adopted the hybrid service. One of the key obstacles is that we lack efficient approaches to find when and where (i.e., zones) we should build the connections between the services with different levels of flexibility (Qiu et al., 2014; Velaga et al., 2012). This study seeks to partially make up this research gap in the literature.

In short, the above review indicates several research gaps that this study tries to make up by developing new methods. First, even though the benefits of coordinating transit systems and emerging on-demand services have been recognized, we still lack efficient approaches to find those critical transfer locations and times for the better coordination. Next, even though various data, including ridership data (i.e., transit smart card data or ridesharing pick/drop data), taxi data, and mobile data, combined with the demographic and land-use features have been extensively studied to explore transit or ridesharing service gaps and design, few studies have investigated crowdsourced trajectory data collected from both transit and ridesharing services. The potential ability of such informative data has not been well explored in the literature. Moreover, the trajectory data involving two mobility service modes are big data presenting non-additive curves spanning in a local network during a time period. Existing approaches, such as various choice models and regression analysis, which have been successfully used to analyze ridership, land use, and demographic data, cannot be directly applied to study the trajectory data. It calls for new data analysis approaches, for which we formally define the problem and introduce our methodology in the following sections.

Problem description

This study is devoted to developing an innovative crowdsourced data analysis approach to help identify potential critical transfer zones, which are proper candidates for coordinating transit routes and on-demand services to promote hybrid mobility services. To do that, we noticed that the ridesharing services are highly correlated with transit services. Accordingly, investigating their trajectory data will help reveal the interaction between these two types of mobility services and further identify the good candidate zones to coordinate them by proper schemes. Along with the above thought, this study considers these trajectory data collected with a general format as follows.

Specifically, this study considers that each trip is formed by a pick-up and drop-off ridesharing service from its origin to its destination and the route it goes through. Similarly, each transit (bus/metro) service running from its first stop to its last stop along its planned route is considered one transit trip. Two transit services occur on the same route, but different schedules are considered two transit trips. Accordingly, we work on V number of ridesharing trips and B number of transit trips. The trajectory of a ridesharing service during each trip v is updated at discrete time stamps $n \in \mathcal{N}_v = \{0,1,...,N_v\}, \forall v \in V$, according to the updating frequency of the ridesharing vehicle's GPS, while the trajectory of a transit during each trip b is updated at discrete time stamps $m \in \mathcal{M}_b = \{0,1,...,M_b\}, \forall b \in B$ corresponding to the transit vehicle arrival time at the stops of the b-th transit trip. Please note that transit and ridesharing trajectory data are often updated with different rates. To mathematically present the trajectory data, we use superscripts $\mathcal{S}$ and $\mathcal{T}$ to differentiate ridesharing and transit trajectories. Accordingly, the trajectory of the ridesharing trip v is denoted as $\mathcal{Z}_v^{\mathcal{S}} = \{\boldsymbol{z}_{\boldsymbol{v},\boldsymbol{n}}^{\boldsymbol{\mathcal{S}}}(\boldsymbol{x},\boldsymbol{y}), n \in \mathcal{N}_v\}, \forall v \in V$, where $\boldsymbol{z}_{\boldsymbol{v},\boldsymbol{n}}^{\boldsymbol{\mathcal{S}}}(\boldsymbol{x},\boldsymbol{y})$, abbreviated as $\boldsymbol{z}_{\boldsymbol{v},\boldsymbol{t_n}}(\boldsymbol{x},\boldsymbol{y})$, is the coordinates of the ridesharing vehicle in trip v at the n-th timestamp (i.e., at time t_n). Particularly, we denote the tuple of $(o_v,d_v), \forall v \in V$ as the origin and destination location of the v-th ridesharing trip. Then, we denote the trajectory of transit trip b as $\mathcal{Z}_b^{\mathcal{T}} = \{\boldsymbol{z}_{\boldsymbol{b},\boldsymbol{m}}^{\boldsymbol{T}}(\boldsymbol{x},\boldsymbol{y}), m \in \mathcal{M}_b\}, \forall b \in B$, where $\boldsymbol{z}_{\boldsymbol{b},\boldsymbol{m}}^{\boldsymbol{T}}(\boldsymbol{x},\boldsymbol{y})$ is the coordinates of the transit at the m-th time stamp in b-th transit trip. We mark t_0 as the departure time at the first stop and $t_m, m \in \mathcal{M}_b$ is the arrival time of the transit at the following stops in b-th transit trip.

Built upon the above trajectory data, this study will develop our data analysis approach, and then conduct a case study to validate the effectiveness of the proposed data analysis approaches. Specifically, the validation involves the analysis of the land-use data, which is the point-of-interest (POI) data indicating whether a location point on the map is commercial/residential or other use types.

Given a very general trajectory data format is considered, our data analysis approaches can be applied to analyze the trajectory data collected from other modes, such as the trip GPS data of bike-sharing services and private vehicles. To facilitate understanding our mathematical formulations, we summarize the important notations in Table 1, even though each of them will be introduced again in the context in the following sections.

Table 1: List of notations

Notation	Explanation
1. Notations in data description	
V	Number of ridesharing trips.
B	Number of transit (bus/metro) trips.
$\mathcal{N}_v = \{n\}_{n=0}^{N_v}$	Set of timestamps that the trajectory of the v-th ridesharing trip is collected.
$\boldsymbol{z}_{v,n}^{S}(\boldsymbol{x},\boldsymbol{y})$	The vehicle location coordinates of the v-th ridesharing trip at time stamp n, abbreviated as $\boldsymbol{z}_{v,t_n}(\boldsymbol{x},\boldsymbol{y})$.
$\mathcal{M}_b = \{m\}_{m=0}^{M_b}$	Set of the transit arrival timestamps at stops of the b-th transit trip.
$\boldsymbol{z}_{b,m}^{T}(\boldsymbol{x},\boldsymbol{y})$	The vehicle location coordinates of the b-th transit trip at time stamp m. (transit stop coordinates)
2. Notations in data presentation and discretization	
T	Data analysis horizon.
$I = \{i\}_{i=0}^{I}$	Set of time intervals by time discretization. I is also used as the maximum number of time intervals.
$K = \{k\}_{k=0}^{K}$	Set of pixels by spatial discretization. K is also used as the maximum number of pixels.
V_i	Set of ridesharing trips during i-th time interval.
τ	Continuous variable represents the length of the time interval.
$\mathbb{Z}_{v,i}^{\tau}$	Set of vehicle location coordinates of the v-th ridesharing trip during the time interval $[i\tau,(i+1)\tau]$ with time interval length τ, $\forall i \in I$.
$\boldsymbol{z}_{v,t}(x,y)$	Vehicle location coordinates of the v-th ridesharing trip at time t.
$\overline{\boldsymbol{z}_{v,\tau}}(\boldsymbol{x},\boldsymbol{y})$	The averaged vehicle location coordinate of the v-th ridesharing trip during the ith time interval with width τ.
$N_{k,i}^{S}$	Number of ridesharing trips that going through the pixel k during the i-th time interval.
$N_{k,i}$	Number of total trips that going through the pixel k during the i-th time interval.
$r_{k,i}$	The ridesharing service ratio of the pixel k during the i-th time interval.
κ	Integer variable represents the total number of the pixels for each time interval.
$u_{k,i}$	Binary variable to identify whether the k-th pixel in i-th time interval is complementary. If true, $u_{k,i} = 1$, otherwise $u_{k,i} = 0$, $\forall k \in K, i \in I$.
$v_{k,i}$	Binary variable to identify whether the k-th pixel in i-th time interval is competitive. If true, $v_{k,i} = 1$, otherwise $v_{k,i} = 0$, $\forall k \in K, i \in I$.
3. Notation in heatmap analysis	
$\mathcal{C}$	Set of clusters (patterns) on a heatmap h.
O_k	Set of the nearest neighbors of a pixel $h(k)$, $\forall k \in K$ on a heatmap h.
$\boldsymbol{h}$	Set of heatmaps. $\boldsymbol{h} = \{h_i, i \in I\}$, in which the heat of the k-th pixel $h(k)$ is r_{ik}.
H	Aggregated heatmap from $\boldsymbol{h}$, in which the heat of the k-th pixel $H(k)$ is $R_k = \frac{1}{K} \sum_{i \in I} r_{k,i}$.
$\hat{\boldsymbol{h}}$	Set of averaged heatmaps for FLM zone analysis. $\hat{\boldsymbol{h}} = \{\hat{h}_\omega, \omega \in \mathcal{W}\}$, in which the heat of the k-th pixel $\hat{h}_\omega(k)$ is $\hat{r}_{k,\omega}$.

$\tilde{\boldsymbol{h}}$	Set of probability heatmaps of FLM-prone orders. $\tilde{\boldsymbol{h}} = \{\tilde{h}_\omega, \omega \in \mathcal{W}\}$, in which the heat of the k-th pixel $\tilde{h}_\omega(k)$ is $\tilde{r}_{k,\omega}$.
o_v	The origin of the v-th ridesharing trip.
d_v	The destination of the v-th ridesharing trip.
γ_k	The number of FLM-prone orders in the k-th pixel of heatmap h.
$\bar{\rho}$	The average density of transit stops over the study region.
V_ω	Set of ridesharing trips that operate during the time interval ω.

Methodology

This section develops our approaches to analyze the trajectory data defined above. Briefly, we first develop an optimal 3D representation approach. It enables us to zoom in and analyze the trajectory intensity and distribution at the discrete level. The analysis results are preented as the heatmaps. Next, we explore the critical transfer zones by aggregating those discrete information from individual heatmap pixels and conducting pattern recognition. These approaches integrate spatiotemporal statistical data analysis, machine learning, and optimization to provide the following capabilities: better understanding the spatiotemporal service correlation of transit and ridesharing on the local transportation network; revealing critical zones (RS zones and FLM zones) for transit and on-demand services integration.

Optimal 3D discrete representation

This study involves ridesharing large-scale GPS trajectory data and transit trajectory data collected with different rates. Specifically, the transit trips have fixed routes and schedules, while the ridesharing trajectories randomly distribute over a city during different time slots over a day. As both temporal and spatial features of the trips are considered, these two sets of data are non-additive and most of the quantitative approaches cannot be directly used to conduct the data analysis. To uncover the service pattern involved in the data, we need a good representation to support the quantitative analysis. Considering the spatiotemporal dynamics of the trajectory data, this study puts all trajectories in a 3D space spanned by 2D (x-y) spatial coordinates and time (t) dimensions. This 3D space represents the entire service space. Accordingly, each trip is presented by a 3D route in the space. See Figure 2(a) for an example, different 3D routes are labeled with different colors.

Furthermore, we notice that these 3D routes intersect and then diverge at different spatiotemporal points. Some areas present very dense trips going through by both transit and ridesharing modes, but others are only sparsely visited by one of them. More importantly, the ridesharing GPS trajectory data are collected at different rates and these trajectory data are not directly addable. To analyze these 3D curves, this study discretizes the 3D service space into K×I number of uniform cubes (see Figure 2(b)), where K is the number of the pixels in spatial area, and I is the number of time intervals in the time dimension. Accordingly, we have K cubes cover a space spanned in a time interval τ over a grid area. With a slight abuse of notation, we also use I to represent the set of time intervals and K to represent a set of cubes (pixels) in one-time interval throughout the paper.

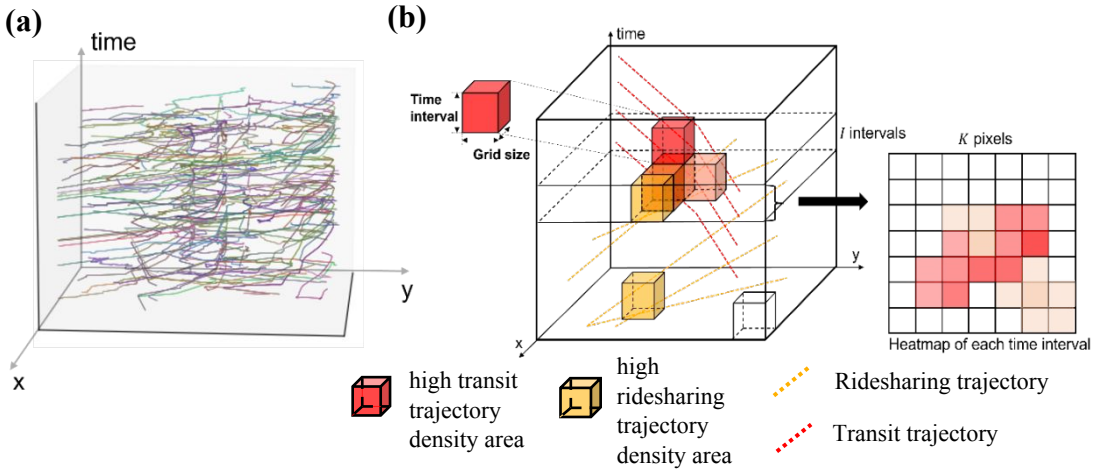

Figure 2. (a) Trips in the 3D space. (b) 3D discretization and heatmap generation.

Once the service space is discretized, we can zoom in and study the service correlation between these two modes in each cube. Specifically, we aggregate the trajectory data of a trip in each cube (see Section: *Optimal length of time interval*), and then we measure the trajectory density of ridesharing mode in a cube by (1) below.

$$R = \{r_{k,i}\}, r_{k,i} = \begin{cases} N^{S}_{k,i}/N_{k,i} & N_{k,i} > 0 \\ -1 & N_{k,i} = 0, \end{cases} \forall k \in K, i \in I, \quad (1)$$

where $N^{S}_{k,i}$ and $N_{k,i}$ denote the number of ridesharing trips and total trips going through the pixel k during the i-th time interval. Equation (1) indicates that $r_{k,i} \in [0,1]$ for $N_{k,i} > 0$ and we mark $r_{k,i} = -1$ if it is an empty pixel (i.e., no trips present in the pixel).

Based on this view, we formally define three types of pixels through the trajectory density defined in (1). Furthermore, we consider those X (transit or ridesharing)-trajectory dominant and trajectory even pixels as informative pixels, as a contrast to the noise pixels defined in Definition 3. Note that both $\underline{\eta}$ and $\underline{\mu}$ should be less than 0.5 to be consistent with the meanings of these definitions.

Definition 1 – X-trajectory Dominant Pixel: A pixel is dominated by X (either ridesharing or transit trajectory), if and only if its trajectory density satisfies $r_{k,i} \in [0,\underline{\eta}] \cup [\overline{\eta},1]$, $k \in K, i \in I$,where $\underline{\eta}$ and $\overline{\eta}$ are given parameters and $\overline{\eta} = 1 - \underline{\eta}$.

An X-trajectory dominant pixel implies that under the current mobility service condition, one service mode within the defined region presents dominant trajectories, which implies this service represents a better or more approachable mobility service than the other one. On the other hand, a trajectory even pixel presents an even mobility supply from both modes. Roughly, we can say they evenly share current demands.

Definition 2 – Trajectory Even Pixel: A pixel presents even trajectory density between the ridesharing and transit if and only if its ridesharing trajectory density satisfies $r_{k,i} \in [\underline{\mu},\overline{\mu}], k \in K, i \in I$, where $\underline{\mu}$, $\overline{\mu}$ are given parameters, $\overline{\mu} = 1 - \underline{\mu}$.

Definition 3 - Noise Pixel: A pixel cannot present a clear trajectory dominant or even relationship between the ridesharing and transit services if its ridesharing trajectory density satisfies $r_{k,i} \in [\underline{\eta},\underline{\mu}] \cup [\overline{\mu},\overline{\eta}]$ (i.e., unidentifiable pixel presenting neither dominant nor even relationship) or $r_{k,i} = -1$ (empty pixel), $k \in K, i \in I$, $\overline{\eta} = 1 - \underline{\eta}$ and $\overline{\mu} = 1 - \underline{\mu}$.

This discretization and definitions enable us to understand the mobility supply of transit and ridesharing at a discrete level. Based on this knowledge, we can further infer when and where the transit and ridesharing can have a good cooperation. From this perspective, it is critical to determine the discretization resolution, e.g., a proper cube dimension (i.e., the length of the time interval and the pixel size in the spatial dimension) to ensure the success of the data analysis. We discuss our ideas in the following sections.

Optimal length of time interval

First of all, we notice that the ridesharing GPS data is collected too frequently (per 2 – 4 seconds), which leads to unnecessary dense data to do the time dimension discretization. The dense data will lead to a huge data size, which results in a significant computation load for future analysis. Therefore, the first task of the

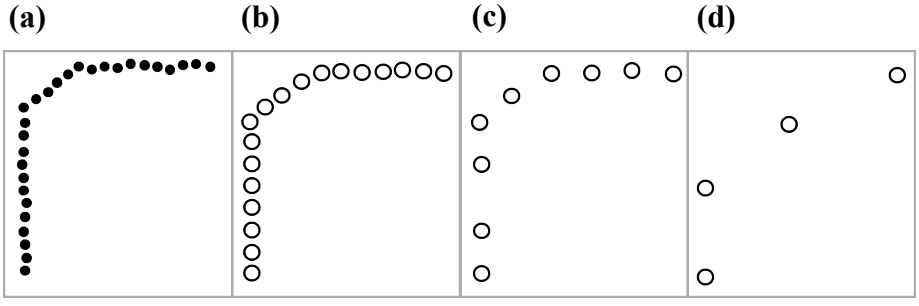

Figure 3. Ridesharing trajectory data aggregation with increasing time interval size from (a) – (b).

discretization in the time dimension is to properly aggregate the data in the time dimension so that we can reduce the computation load without over sacrificing the data resolution. This section investigates how the

time interval of the cubes will affect the statistical analysis and then explores the optimal time interval τ^* for the cubes.

Please note that with a given time interval, we locate the spatial position of a ridesharing vehicle by averaging the coordinates of multiple data points during each time interval so that we can reduce the data size. This process will compromise the accuracy of the spatial information if the vehicle trajectory is oversimplified. Take Figure 3 for example, the actual trajectory of a ridesharing vehicle may go across a wide spatial area (see Figure 3 (a)). However, if the time horizon is too small such as Figure 3 (b), even though the trajectory is well preserved, the temporal discretization leads to many time intervals with many data points. Then, the discretization results in a data set that causes an expensive computation load for the data analysis and machine learning training process. If the time horizon is too large such as Figure 3(d), then the aggregated data do not span over the space as the actual trajectory does. As a result, the resolution of the data is sacrificed. On the other hand, Figure 3 (c) presents a good time interval. It reduces the raw data size while keeping the same spatial coverage. In other words, this discretization reduces the computation load in data analysis without over sacrificing the data resolution. Therefore, we prefer a selection of time interval τ, which balances the computation load and information accuracy. By recognizing this point, we develop the optimization model in (2)-(5) to optimally aggregate the ridesharing trajectory data collected in the analysis horizon T. More exactly, it explores the optimal time interval with the objective to minimize the information loss and the dataset size, subject to a feasible range of the value τ. Accordingly, the objective function uses the sample variances (the first item in (2)) to measure the loss of the location information and puts the penalty on the number of the time intervals (the second item in (2)).

$$\boldsymbol{P_1}: \min_{\tau} \alpha \sum_{i=0}^{I} \frac{1}{|V_i|} \sum_{v \in V_i} \frac{1}{|\mathbb{Z}_{v,i}^{\tau}|} \sum_{i\tau \le t \le (i+1)\tau} \left|\left| \boldsymbol{z_{v,t}}(\boldsymbol{x},\boldsymbol{y}) - \overline{\boldsymbol{z_{v,\tau}}}(x,y) \right|\right|_2^2 + \beta \frac{T}{\tau} \tag{2}$$

Subject to,

$$\underline{\tau} \le \tau \le \bar{\tau}, \tag{3}$$

$$I = \lfloor T/\tau \rfloor, \tag{4}$$

$$\overline{\boldsymbol{z_{v,\tau}}}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{|\mathbb{Z}_{v,i}^{\tau}|} \sum_{i\tau \le t \le (i+1)\tau} \boldsymbol{z_{v,t}}(x,y), \forall v \in V_i, i \in I \tag{5}$$

where τ, the length of the time interval, is the single decision variable and $\boldsymbol{z_{v,t}}(x,y)$ is the input data, which represents the vehicle location coordinates of the v-th ridesharing trip at time t; $i \in I$ is the index for time interval; I in constraint (4) also represents the total number of time intervals, which changes according to the decision variable τ; V_i represents the set of ridesharing trips during i-th interval; $\mathbb{Z}_{v,i}^{\tau} \subset \mathcal{Z}_v^S$ is the set of vehicle coordinate records of v-th ridesharing trip during the time interval $[i\tau,(i+1)\tau]$; specifically, $\mathbb{Z}_{v,i}^{\tau} = \{\boldsymbol{z_{v,t}}(x,y) \,|\, i\tau \le t \le (i+1)\tau\}$; T is the analysis horizon; α and β are the predefined weights normalized to make the two terms comparable in the objective function in the magnitude; $\overline{\boldsymbol{z_{v,\tau}}}(\boldsymbol{x},\boldsymbol{y})$ is the averaged vehicle location coordinates of the v-th ridesharing trip during the i-th time interval with width τ (constraint (5)). The optimization Model $\boldsymbol{P_1}$ aims to find an optimal time discretization so that we can balance the data resolution and the computation load. The objective function (2) factors the data variance missing (i.e., referring to resolution loss) introduced by averaging the ridesharing trajectory coordinates and the number of time intervals resulting from the discretization (i.e., referring to the computation load). When the time interval length τ is small, the location data of the vehicle $\boldsymbol{z_{v,t}}(x,y)$ do not change much during the time interval so that the average coordinates $\overline{\boldsymbol{z_{v,\tau}}}(\boldsymbol{x},\boldsymbol{y})$ well represent the location of the vehicle, see (5). Accordingly, the data variance missing as the first term of (2) is small (i.e., good data resolution) but a small τ leads to many time intervals and makes the second term of (2) large (i.e., high computation load in data analysis). On the other hand, when τ is too large, the trajectory is oversimplified so that the second term is small (i.e., low computation load in data analysis) but with large variance missing as a tradeoff (i.e., bad data resolution).

Even though model P1 presents nonconvexity (the decision variable presents as the upper bound of the summation), it only involves one decision variable bounded by a box constraint. Thus, we can find a lower and upper bound for τ in practice to limit the solution space (see more discussion in the case study). And then, we can quickly search the local optimal solution of τ by using the best first search (BFS) algorithm (Dechter & Pearl, 1985) in a feasible region $[\underline{\tau},\bar{\tau}]$ (constraint (4)).

BFS is one of the efficient sequential search algorithms in discrete optimization. It is a selected enumerated method. For completeness, we present the main idea of this algorithm as follows. The BFS maintains a list named OPEN, which is placed with nodes (possible solutions) to be expanded. Initially, the list of OPEN includes a set of integer solutions within $[\underline{\tau},\bar{\tau}]$. Then, the solutions are evaluated through the objective function (2). The worst solution is removed from the list and the best solution is expanded to

include its neighbors in the OPEN list as successors. The heuristic evaluation process is repeated until no more successors are found. The best solution that remains in the OPEN list is the optimal solution.

Note that the transit trajectory data is only collected at each stop. Therefore, it is not necessary to do further aggregation in the time dimension, even though model $\boldsymbol{P_1}$ can be applied to transit trajectory data if they are very dense.

Optimal pixel size

This study next determines the pixel size of the cubes in the spatial dimension. Built upon the aggregated trajectory data, the spatial discretization aims to maximize the number of informative cubes in each time interval so that we can accurately uncover valuable information from the data. To do that, we notice that the size of the pixel will also affect the number of the cubes as well as the counting of the ridesharing/transit trips which is closely relevant to the trajectory density measurement occurring in each cube, see (1). Thus, it will influence the computation load as well as the interpretation of the data analysis. We use Figure 4 as an example to explain this point. We consider that the average speed of the overall traffic data is v_a and the average speed upper limit over the network is $\overline{v}$. With the given time interval width τ, if the size of the pixel is too small, such as the length of the pixel side $l < \tau v_a$ in Figure 4 (a), it will very likely lead to many empty pixels since the majority of vehicles can run across a pixel during a time interval τ. On the other hand, if the size of the pixel is too large, such as the length of the pixel side $l > \tau\overline{v}$ in Figure 4 (b), it will lead to the overcount since an individual vehicle will be caught more than once in the pixel.

The example above shows that a pixel with the length within $[\tau v_a, \tau\overline{v}]$ will facilitate the analysis better. An improper discretization will lead to either many empty cubes or overcount. It cannot provide valuable insights and will affect the statistical analysis significantly. Therefore, we seek to explore the optimal pixel size so that they can clearly present either trajectory dominant or even relationship between traffic modes. According to Definitions 1 and 2, an X-trajectory dominant pixel indicates that either the transit or the ridesharing trajectory dominates the service going through the pixel. Potentially, many adjacent and connected trajectory dominant pixels show one out of the two services dominates in this area. On the other

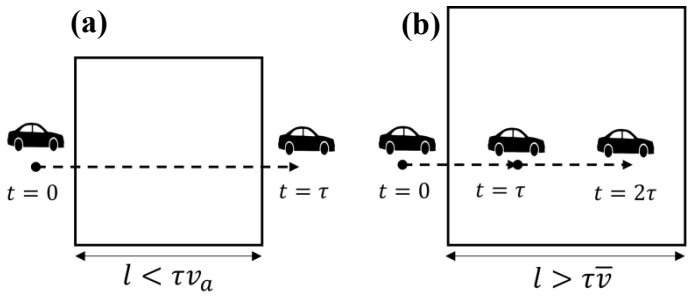

Figure 4. Examples of improper pixel size. (a) $l < \tau v_a$, (b) when $l > \tau\overline{v}$.

hand, an area formed by connected even pixels shows that neither the transit nor the ridesharing presents apparent merits to attract more the trips going through it. This information is very valuable to facilitate our data analysis later. Clearly, the ratios of informative pixels (X-trajectory dominant and even) and noise pixels are affected by the schemes of the 3D discretization in the spatial dimension. We next discuss our ideas to search for the optimal discretization scheme in spatial dimension by an optimization model formed in (6)-(9). It seeks to find the optimal number of pixels (κ^*), so does the size of each pixel, for maximizing the total number of the informative pixels (X-trajectory dominant and even) with a given time interval width τ. The objective function (6) consists of $A_{\kappa,i}$ and $\overline{A}_{\kappa,i}$, which are the proportion of the X-trajectory dominant and even pixels over the study region; p and q are predefined weights.

$$\boldsymbol{P_2}\text{: } \max_{\kappa} \sum_{i=0}^{I} (pA_{\kappa,i} + q\overline{A}_{\kappa,i}) \tag{6}$$

Subject to,

$$A_{\kappa,i} = \frac{1}{\kappa}\sum_{k=1}^{\kappa} u_{k,i}(\kappa),\ \forall i \in I \tag{7}$$

$$\overline{A}_{\kappa,i} = \frac{1}{\kappa}\sum_{k=1}^{\kappa} v_{k,i}(\kappa),\ \forall i \in I \tag{8}$$

$$\underline{\kappa} \leqslant \kappa \leqslant \overline{\kappa} \tag{9}$$

In the model $\boldsymbol{P_2}$, κ is the decision variable, representing the total number of the pixels, each with a square shape; $u_{k,i}$ and $v_{\kappa,i}$ are auxiliary binary variables, $u_{k,i} = 1$ if the k-th pixel in i-th time interval is X-

trajectory dominant with $r_{k,i} \in [0,\underline{\eta}] \cup [\overline{\eta},1]$, and 0, otherwise; $v_{k,i} = 1$ if the k-th pixel in i-th time interval has even trajectory density with $r_{k,i} \in [\underline{\mu},\overline{\mu}]$, and 0, otherwise; p and q are predefined weights. Constraints (7) and (8) count the proportion of the X-trajectory dominant pixels ($A_{\kappa,i}$) and the even pixels ($\overline{A}_{\kappa,i}$). Therefore, the objective function (6) seeks to maximize the proportion of the X-trajectory dominant and even pixels over the study region. Note that for a given study area with the size S, the more pixels present, the smaller the pixel size is. As a proper pixel length is within $[v_a\tau,\overline{v}\tau]$, the total number of pixels, κ, is also bounded by $[\underline{\kappa},\overline{\kappa}]$ (constraint (9)), where $\underline{\kappa} = [S/(\overline{v}\tau)]^2$ and $\overline{\kappa} = [S/(v_a\tau)]^2$.

Next, given $\overline{\eta} = 1 - \underline{\eta}$ and $\overline{\mu} = 1 - \underline{\mu}$ defined in Definitions 1-3, we notice that two of the four parameters (e.g. $\underline{\mu}$ and $\underline{\eta}$) will significantly affect the solution of the discretization since they affect the value of the trajectory density ($r_{k,i}$) used in Definitions 1-3 so do the values of $A_{\kappa,i}$ and $\overline{A}_{\kappa,i}$ in the optimization model. More exactly, this study names the interval $\underline{\mu} - \underline{\eta}$ as an unidentifiable interval (UI). A narrow UI loosens the criteria to certify X-trajectory dominant or even pixels according to Definitions 1 and 2. Accordingly, it tends to produce a discretization solution with a few pixels (i.e., a small value of κ^*) each with a large size, which may maximize the size/number of the informative pixels but result in a low resolution, e.g., a pixel may cover some areas not presenting consistent lane-use features. On the other hand, a wide UI tightens the criteria and leads to a discretization solution with plenty of pixels (i.e., a large value of κ^*) each with a small pixel size, which may lead to more noise pixels as the cost. The determination of the parameters is highly data orientated. Thus, this study will perform a sensitivity analysis for the UI in our case study. Combining with the land-use analysis, we suggest proper values for the parameters $\underline{\eta}$, $\overline{\eta}$, $\underline{\mu}$ and $\overline{\mu}$ in the case study.

The optimization model $\boldsymbol{P_2}$ is nonlinear and nonconvex, but with a single integer decision variable κ within $[\underline{\kappa},\overline{\kappa}]$. This study thus explores the optimal solution by using the similar best-first search (BFS) heuristic approach (Dechter & Pearl, 1985), which is also used in solving the model $\boldsymbol{P_1}$.

Remark: The discretization in the time dimension mainly seeks to aggregate the dense trajectory data, while balancing the data resolution and computation loads. The spatial discretization aims to maximize informative cells, while balancing the computation load and data interpretation capability. In general, the temporal and spatial discretization should be simultaneously optimized. However, the trajectory data used in this study are collected very frequently (per 4 seconds), which is not necessary for our data analysis purpose. Accordingly, the data aggregation in the temporal dimension does not affect the spatial discretization very much, but this is not true in the other direction. Therefore, our 3D discretization approach will first determine the optimal time interval to aggregate trajectory data. Built upon that, we decide the optimal pixel size to generate sufficient informative cubes for further data analysis.

Heatmap generation and pattern recognition

Built upon the 3D discretization, we can capture the trajectory distribution feature in each cube. More exactly, we present the trajectory density of ridesharing/transit in each cube during one time interval τ^* by

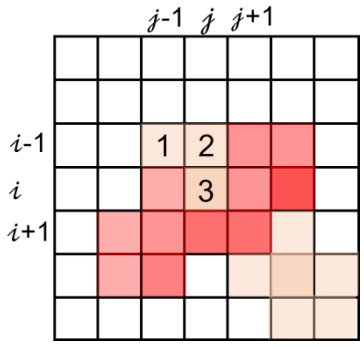

Figure 5. An example of heatmap and clustering.

a heatmap, in which the heat of each pixel represents its ridesharing trajectory density. See the example shown in Figure 2, the color of each pixel demonstrates the intensity of the ridesharing/transit trajectories. A light-yellow pixel corresponds to a cube with high ridesharing trajectory density, while the dark-red pixel indicates a cube mainly gone through by transit trajectories. Applying this analysis to each interval in the 3D space, we obtain a set of heatmaps $\boldsymbol{h} = \{h_i,\ i \in I\}$. However, these analysis results are information fragments for individual cubes. They can neither demonstrate the service patterns with collective features

nor infer critical zones over the spatiotemporal space of interests. To solve this issue, we next seek to recognize the critical patterns with collective features on a heatmap so that we can discover the converged insights. That is to find collective pixel clusters according to their colors. For example, the yellow zone involving pixels 1, 2, and 3 in Figure 5 together indicates a ridesharing trajectory dominant zone, while the red zone represents a transit trajectory dominant zone with a low ridesharing trajectory density. By analyzing these clusters, we seek to discover critical transfer zones, which suggest service gaps and indicate the need of integrating transit, microtransit, and ridesharing services.

To conduct this pattern recognition, our main idea is to merge the pixels into a set of clusters $\mathcal{C}$ on a heatmap h (for generality, we omit the subscript i here) according to the given clustering criterion. The pixels within a cluster, such as $C \in \mathcal{C}$, should exhibit the same service property, i.e., similar color, ensured by the clustering criterion. For example, we can gather the pixels to a cluster if the heat or the pixel's ridesharing trajectory density, r_k satisfies $r_k \geq r, \forall h(k) \in C$, where r is a given threshold value. When the clustering criterion is $r_k \geq r$ and r is given as a big value, i.e., $r = 0.9$, the cluster that we try to recognize represents a ridesharing trajectory dominant zone. On the other hand, when the clustering criterion is $r_k \leq r$ and r is a small value, i.e., 0.1, the cluster that we try to recognize is a transit trajectory dominant zone. In other words, these clusters exhibit collective and converged features regarding the relative mobility supply of transit and ridesharing services.

We present the clustering approach by Algorithm 1. It borrows the idea of the k-nearest neighbors algorithm (KNN), but with our development according to the problem features. Specifically, we consider all direct adjacent pixels, denoted by O_k, as the nearest neighbor set of a pixel $h(k), \forall k \in K$ be located by the index (i,j) in the heatmap. Accordingly, the set O_k involves the eight pixels with index $(i-1,j-1)$, $(i-1,j)$, $(i+1,j)$, $(i,j$-1), $(i,j$+1), $(i+1,j-1)$, $(i+1,j)$, and $(i+1,j+1)$ (see Figure 5). The algorithm starts with all candidate pixel seeds, each of which satisfies $r_k \geq r$ (or $r_k \leq r$), $\forall k \in K$, and then repeatedly has each seed clustering its nearest neighbors into one cluster according to the clustering criterion, until all pixels are merged into clusters. The pseudocode of this clustering algorithm is presented in Algorithm 1. It takes a heatmap h and a given threshold r as inputs, and returns the cluster set $\mathcal{C}$. The steps in line 3 – line 5 form a list of the candidate pixel seeds (*CPL*) by the given clustering criterion. The steps from line 7 – line 13 cluster each pixel $h(k)$ in *CPL* with its nearest neighbor pixel $h(l) \in O_k$ by the clustering criteria. The nearest neighbor pixels set of $h(k)$ is updated with $O_k \cup O_l$ and $h(l)$ is popped out from *CPL*. The process is repeated until no more nearest neighbor pixel $h(l)$ can be clustered (line 7 – line 13).

Algorithm 1 Clusters Construction

```
1   Procedure CLUSTER(h, r)
2       CPL ← ∅
3       for pixel h(k) ∈ h do
4           if r_k ≥ r then  # clustering criterion, can also be r_k ≤ r
5               CPL ← k
6       C_k←k, ∀k ∈ CPL; C = {C_k}
7       for k ∈ CPL do
8           for pixel h(l) ∈ O_k do
9               if l ∈ CPL and l ∉ C_k then # h(l) meets clustering criterion, r_l ≥ r
10                  C_k←C_k ∪ C_l
11                  O_k←O_k ∪ O_l
12                  C_l←∅
13                  CPL ← CPL\{l}
14      return C
```

Applying the clustering algorithm to the heatmaps, we can recognize the service pattern over the study area during each time interval. This information further enables this study to conduct two specific heatmap-based analyses, which are introduced in the following two sections.

Searching Ridesharing Swarm Zones

We are first interested in a critical pattern: the ridesharing swarm (RS) zones in the heatmaps, which

are the zones spatiotemporally passed by a crowd of ridesharing trips, but not transit trips. Given it is a new concept introduced by this study, this section first conceptually discusses the importance of RS zones. Then, we introduce the approach to recognize RS zones on a heatmap. Later, our numerical will validate our thoughts.

Using Figure 6 as an illustration example, the RS zones can be the zones where either many ridesharing picked-up or dropped-off services happen, or many ridesharing trips pass by. In either situation, this study considers that these RS zones indicate promising critical transfer zones (see Figure 1), which most likely lack transit connectivity and are good candidate zones for integrating the transit and on-demand services. Below we provide a further discussion regarding this correlation and the importance of RS zones.

First, if there are consistent and dense ridesharing trajectories going through a pair of RS zones (please note that a zone here will cover a large area with multiple roads), this pair of RS zones indicates a potential corridor zone that has a hidden transit connection gap/need, even though it may not indicate complete routes. Adding new transit stops (or routes) in (or between) RS zones can potentially improve the connectivity of these zones to the nearby transit network and then attract more potential transit usages. Next, if an RS zone does not have sufficient passing trips with other RS zones in the map, but covers a large area surrounded by transit trajectory dominant zones in a heatmap, it indicates a good place to setup transfer connections by either microtransit or other types of on-demand services for connecting this zone to nearby transit service. It will also improve transit connectivity and potentially initiate more transit demand. Our case study confirms these thoughts above by the field examples. Both scenarios above indicate that RS zones imply promising critical zones, even though extra data and sophisticated operation studies are needed to further determine the operation decisions, which are out of the scope of this study.

Motivated by the above thoughts, this study develops an RS zone recognition approach upon the heatmaps generated from the ridesharing trajectory data. We introduce it as follows. Given RS zone shows valuable aggregated spatiotemporal insights, we project all heatmaps $h_i \in \boldsymbol{h}$ in the study to the spatial region to obtain a new aggregated heatmap H. The heat R_k of the pixel $H(k)$ is calculated by averaging the heat $r_{k,i}$ over all heatmap $h_i \in \boldsymbol{h}$, see (10).

$$R_k = \frac{1}{K}\sum_i r_{k,i},\ \forall i \in I, k \in K \tag{10}$$

The heatmap H presents the overall service information from the study horizon, i.e., one month in the case study. We then demonstrate the RS zones recognition and its capability to uncover the hidden transfer zones with good potential to integrate transit and on-demand services. Specifically, we take $(H, \bar{\eta})$ as the

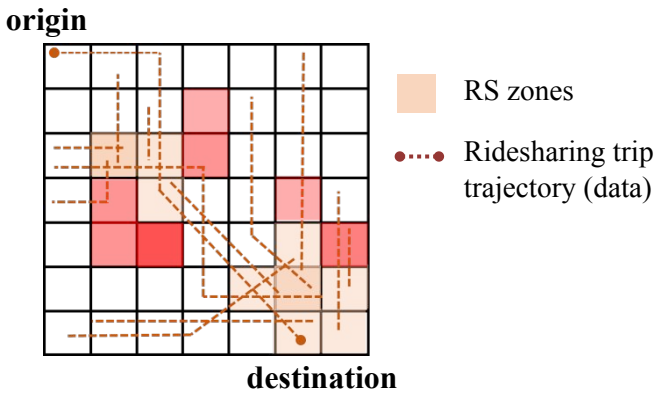

Figure 6. RS zones on heatmap H.

input and recognize the RS zones (clusters) by using the algorithm *CLUSTER*$(H, \bar{\eta})$, using the clustering criterion, $R_k \geq \bar{\eta}, \forall H(k) \in C$. The threshold parameter $\bar{\eta}$ is selected according to Definition 1 in Section: *Optimal 3D discrete representation*. Mainly, the ridesharing trajectory is dominant within the area with $R_k \geqslant \bar{\eta}$. These RS zones are closely correlated to potential connection zones for integrating transit and on-demand services. Our case study in Section: *Heatmap generation and pattern recognition* validates the conjecture by a case study combining land-use analysis. Note that the RS zones can only be recognized by analyzing the entire trip (trajectory) data, which highlights the value of trajectory data analysis in this study.

Searching First and Last Mile (FLM) Gap

Suffering from the limited coverage and flexibility in the current transit system, transit passengers often

meet the difficulty of the first and/or last mile (FLM) gaps. They then switch to ridesharing or private auto modes to obtain more convenient mobility services. Emerging on-demand services (such as microtransit) provide a promising solution to make up for this deficiency. However, it is very hard to find the first and last-mile gaps due to the lack of intermodal trips or relevant survey data. This study develops a new approach to infer the candidate first/last mile zones through heatmap analysis. Specifically, we notice that those "sandwich" patterns on the heatmaps, in which a transit trajectory dominant zone is immediately connected by multiple ridesharing trajectory dominant zones (see Figure 7, i.e., A_1BA_2), have a great potential to be the zones presenting first/last mile gaps.

To conceptually interpret this motivation above, we consider a large area, such as a big shopping center or residential area in reality. If there are transit stops nearby this area, the first/last mile gaps often happen if the transit stops and the shopping center are beyond walking distance. Accordingly, we are expected to see many ridesharing services there. We noticed that this scenario, in reality, is highly correlated to the "sandwich" patterns we defined above. We next present the correlation between the FLM zones and the "sandwich" pattern on the heatmaps, combining the transit stops distribution data. Taking Figure 7 as an example, the zones with yellow color (such as A_1, A_2 zones) represent the ridesharing trajectory dominant zones (i.e., $\hat{r}_{k,\omega} \geq \overline{\eta}$). The zones in red color (such as B zones) are transit trajectory dominant zones ($\hat{r}_{k,\omega} < \underline{\eta}$). Then, $A_1 - B - A_2$ forms a "sandwich" pattern of our interests. If the transit stops density $\rho(.)$ of the zones A_1, A_2 and B satisfy the relation $\rho(B) > \rho(A_1)$, $\rho(B) > \rho(A_2)$, we know that both A_1 and A_2 have a low transit coverage inside, just like the big shopping center, but plenty of transit stops nearby connecting to B zone. If we also observe many trips between A_1 to A_2 (i.e., $A_1 \leftrightarrows A_2$ ridesharing trip) using ridesharing not nearby transit services in zone B, we conjecture that FLM gaps will most likely occur at A_1 and A_2 zones. On the other hand, if we observe many ridesharing trips between A zone and B zone: $A_1 \leftrightarrows B; A_2 \leftrightarrows B$. It also implies that passengers in A_1 and A_2 zones may meet the difficulty to approach the transit services in zone B. In either case, we can see that A zones in the "sandwich" patterns demonstrate a strong correlation to the potential FLM zones. Our case study confirms this thought by examining field data (see Section: *Searching RS zones*).

Invoked by the above observation, we next formally develop our approach to search FLM zones by recognizing those "sandwich" patterns on the heatmaps. To do that, we first process our heatmaps. Given the FLM demand usually varies from hour to hour (Shen et al., 2018), our analysis is built upon a set of hourly aggregated heatmaps $\hat{h}_\omega \in \hat{\boldsymbol{h}}$, where $\mathcal{W} = \{1,2,\ldots,\left\lfloor \frac{I}{n\tau} \right\rfloor\}$, each heatmap $\hat{h}_\omega$ is aggregated from n number of heatmaps $h_i \in \boldsymbol{h}$, corresponding to one hour time period. We also denote $V_\omega \subset V$ as a set of ridesharing trips that operate during the time range on a heatmap $\hat{h}_\omega$. According to the heatmaps $\hat{\boldsymbol{h}} = \{\hat{h}_\omega, \omega \in \mathcal{W}\}$, we define another set of heatmaps, $\tilde{\boldsymbol{h}} = \{\tilde{h}_\omega, \omega \in \mathcal{W}\}$, each of which has the same interval length as $\hat{h}_\omega$, but the heat of each pixel represents the FLM probability during time interval ω. In other

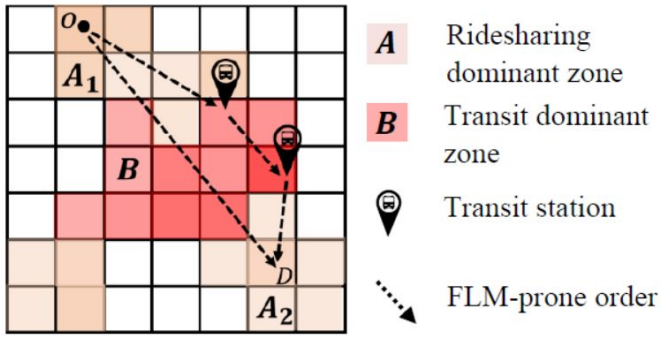

Figure 7. Schematic representation of the "sandwich" pattern.

words, those areas with high heat on the heatmaps $\tilde{\boldsymbol{h}}$ indicate promising FLM zones. To construct such heatmaps $\tilde{\boldsymbol{h}}$, we propose an FLM zones search algorithm (Algorithm 2), which takes the averaged heatmaps $\hat{\boldsymbol{h}}$ as input and provides the FLM probability heatmap $\tilde{\boldsymbol{h}}$. The pseudocode can be found in the Appendix.

Mainly, Algorithm 2 (line 2 to line 3) first applies the clustering algorithm (Algorithm 1) to the heatmap $\hat{h}_\omega$ to find a set of ridesharing trajectory dominant zones $\mathcal{A}$ (with ridesharing trajectory density greater than $\overline{\eta}$) and transit trajectory dominant zones $\mathcal{B}$ (with ridesharing trajectory density less than $\underline{\eta}$). Then the average density of transit stops over the study region, $\overline{\rho}$ is used as a threshold value to filter out undesired zones from $\mathcal{A}$ and $\mathcal{B}$ (line 4 to line 9). For example, a potential FLM zone $A_k \in \mathcal{A}$ should have a transit service coverage lower than the average density, while a transit trajectory dominant zone $B_k \in \mathcal{B}$ will have sufficient transit services (greater than the average density). Next, the algorithm recognizes the "sandwich"

patterns (line 10 to line 15). Specifically, for each $B_k \in \mathcal{B}$, the algorithm finds all ridesharing trajectory dominant zones $A_k \in \mathcal{A}$ that are connected with B_k and puts them in a list B_k^A. Following that, we locate the FLM zones on the "sandwich" patterns by examining the number of FLM-prone orders, γ_k occurred for each pixel within ridesharing trajectory dominant zones $A_k \in \mathcal{A}$. Accordingly, three types of FLM-prone orders are considered, of which trips are from A_k to A_l without using transit (line 16 to line 23), from A_k to the nearby areas in B_k to cover the first-mile with ridesharing (line 24 to line 29), and from pixels in B_k boundary to A_k to cover the last-mile with ridesharing (line 30 to line 34). Finally, the algorithm returns an FLM-prone order probability heatmap $\tilde{h}_\omega$ with the heat $\tilde{r}_{k,\omega}$ as the proportion of the FLM-prone orders that pixel $\tilde{h}_\omega(k)$ receives, $\tilde{r}_{k,\omega} = \gamma_k / \sum_k \gamma_k$.

With Algorithm 2, the heatmaps from $\hat{\boldsymbol{h}} = \{\hat{h}_\omega, \omega \in \mathcal{W}\}$ can be fed into *FLMZONES*($\hat{h}_\omega$, $\bar{p}$) to generate the set of FLM-prone order probability heatmaps $\tilde{\boldsymbol{h}} = \{\tilde{h}_\omega, \omega \in \mathcal{W}\}$. These two sets of heatmaps are input into a two-channel ConvLSTM model for prediction in the next section. The proposed approach identifies the potential FLM zones spatiotemporally, which allows transit agencies to efficiently incorporate flexible services, e.g. as a referenced FLM demand radar for microtransit service planning. Please note that this approach is to scope potential FLM zones. Combining more data into further investigation and validation, such as multimodal trip data and land use data, we can sharpen the results further. Our case study in Section: *Inferring FLM zones* will validate our findings by combining transit stop density and land-use analysis.

Learning Spatiotemporal Service Gaps

The above sections manage to discover the hidden transfer zones for integrating transit and on-demand mobility services, based upon aggregated heatmaps along with a given timeframe. They provide limited help for flexibly responding to demand variation over different timeframes since the temporal FLM demand pattern are not learned. This study is thus inspired to develop the deep learning model to learn and predict those service gaps of interest.

Specifically, we employ a two-channel ConvLSTM learning model using the heatmap data, i.e., $\hat{h}_\omega \in \hat{\boldsymbol{h}}$ and $\tilde{h}_\omega \in \tilde{\boldsymbol{h}}$ as inputs to predict the FLM zones. We justify the selection of this model by the reasons as follows. One of the key characteristics of these data is the high spatiotemporal correlation. For example, some areas on the heatmap $\hat{h}_\omega$ may share common features, such as the areas near the metro hub attract significant mobility needs, including ridesharing demand or/and FLM demand. Moreover, the heatmaps, such as $\hat{\boldsymbol{h}}$ and $\tilde{\boldsymbol{h}}$, are the sets of time series data. It's important to capture this spatiotemporal correlation in the prediction model to improve accuracy. Recent advances in deep learning have enabled researchers to model complex nonlinear relationships. More exactly, a Convolutional neural network (CNN) has been used to capture complex spatial correlation (J. Zhang et al., 2016), and Long Short Term Memory network (LSTM) has exhibited outstanding performance on time series data prediction. The ConvLSTM model, a combination of CNN and LSTM (Xingjian et al., 2015), has demonstrated the satisfying performance to capture the spatiotemporal correlation in the data for weather precipitation forecast prediction. Given the spatiotemporal characteristics of heatmap data, this study, therefore, uses the ConvLSTM model (Xingjian et al., 2015) to predict the dynamics of transit service gaps.

More exactly, we adopt a two-channel ConvLSTM model, with the first channel as the heatmap data $\hat{h}_\omega \in \hat{\boldsymbol{h}}$ and the second channel as the FLM-prone order probability heatmap data, $\tilde{h}_\omega \in \tilde{\boldsymbol{h}}$. Recall that the FLM heatmap, $\tilde{\boldsymbol{h}}$, is generated upon the analysis of OD information and heatmaps $\hat{\boldsymbol{h}}$. Therefore, the heatmaps data, $\tilde{\boldsymbol{h}}$ and $\hat{\boldsymbol{h}}$, are correlated with each other. This two-channel build-up maintains the correlations between different channel data, which can accurately and simultaneously predict the heatmaps for analysing the ridesharing swarms and FLM zones. Specifically, the input data is a sequence of S $\times$ S, two-channel images, or a series of tensors, $X_t \epsilon R^{S \times S \times 2}$. We validate the performance of the two-channel ConvLSTM learning model by the case study below.

Case study

This case study validates the capability of the proposed crowdsourced data analysis approach to find the critical transfer zones for transit and ridesharing cooperation based on the field trajectory data. More exactly, we will examine if the discretization approach can efficiently support the data analysis and further validate whether critical zones we found can infer high potential needs for cooperating transit with on-demand services. Please note that the hybrid mobility service is an emerging field and there are not consistent/well-accepted criteria in the literature to determine if a candidate zone must need such cooperation. It is another research gap and does not obtain enough attention in the literature yet. Therefore,

we validate our approaches and results according to the land use and limited literature. The validation focuses on showing the high possibility of correctness but not an exact yes-or-no answer.

The case study is built upon the testbed consisting of the transit and ridesharing services data in the city of Chengdu. It is a major city in China and has a population of 7.8 million. As shown in Figure 8, the study area is in the second ring region of the city and covers a square region with 5 miles edge length. The ridesharing service data is provided by DiDiChuxing Gaia open dataset (https://gaia.didichuxing.com). It involves about 0.2 million trips made by DiDi ridesharing services per day from November 1 to November 30, 2016. The profile of the data includes the GPS trajectory data of ridesharing vehicles, which are updated 2 – 4 seconds, and the ridesharing order requests, which record the pick-up and drop-off timestamps and locations. The public transit data are collected by the website of the Moovit app (https://moovitapp.com/). The data cover all bus lines and subway lines information in Chengdu city. For each line, the profile of the data includes stop names, stop locations, operation times, and transit vehicles' arrival times at each stop. Please note that the real-time trajectory data of transit vehicles are not available in this study. We will use the schedule and stop location data to format each trip. The operating time of transit services varies from line to line, but most of them start from 6 am and end at 8 pm – 10 pm. There are a total of 1226 transit stops distributed within the study area and in total and 246 transit lines passing through the study area. As the study area is $5 \times 5\ mi^2$, there are about 7 transit stops per mile on average and the average transit stop spacing is 880 ft. The case study is run on a DELL Precision 3630 Tower with 3.60GHz of Intel Core i9-9900k CPU and 16 GB RAM in a Windows environment. The following sections introduce the case study for using the data analysis approach developed by the above effort.

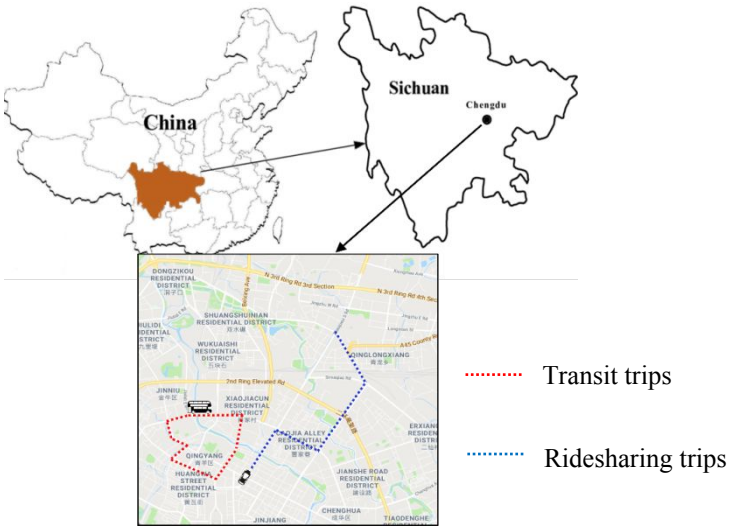

Figure 8. The study area.

Establishing 3D Discretization

We first establish the 3D discretization (τ^*, κ^*) for the ridesharing trajectory data involved in this case study. To do that, we first determine the optimal time interval τ^* by solving the model $\boldsymbol{P_1}$. According to the DiDi data collection rate of 2~4 seconds, we set $\underline{\tau} = 4s$. And then, we make $\overline{\tau}$ equal to the average DiDi single trip service time provided. Accordingly, the initial solution in the OPEN set as [4sec, 1min, 2min, 5min, 10min, $\overline{\tau}$]. Even though $\overline{\tau}$ is a large upper bound, the exploration noticed that the objective value of P1 keeps increasing once $\tau \geqslant$ 5min. This is because the increment of the information loss dominates the dataset size reduction. Given $\boldsymbol{P_1}$ is a minimization model, the search algorithm will discharge the solutions $\tau \geqslant$ 5min quickly, and reduce the search space to [4sec, 2min] only after 3 iterations. It takes the program $\boldsymbol{P_1}$ about 1100s to obtain the solution $\tau^* = 90s$, with which we average about 30 coordinates of each vehicle to locate it in a time interval. Next, we explore the optimal κ^* through the program $\boldsymbol{P_2}$, which needs to pre-determine the parameters $(\underline{\eta}, \underline{\mu})$. By setting $\underline{\eta} = 0.1, \overline{\eta} = 1 - \underline{\eta} = 0.9, \underline{\mu} = 0.4, \overline{\mu} = 1 - \overline{\mu} = 0.6$, it takes the program $\boldsymbol{P_2}$ around 2500 seconds to find the optimal number of the pixels, $\kappa^* = 28^2$, which indicates a pixel size: 325m $\times$ 325m for each cube.

Sensitivity analysis

According to our discussion in Section: *Optimal pixel size*, we justify the selection of the parameters η, and μ for this case study by doing the sensitivity analysis on the length of the UI, i.e., the length of $(\mu - \eta)$. Mainly, with a given time interval τ^*, we test the performance of the 3D discretization under each of the four UIs shown in Table 2, where the UI varies from 0.1 to 0.4 and each corresponds to a set of parameters selection.

Program $\boldsymbol{P_2}$ is run under each UI and generates the optimal solution κ^* shown in Table 2. Upon each optimal discretization scheme (τ^*,κ^*), the heatmaps $\boldsymbol{h} = \{h_i, i \in I\}$ are generated and shown as the examples in Figure 9, in which the region completely dominated by transit or ridesharing trajectory is colored red and white respectively, but the regions where neither transit nor ridesharing services show are in black. Accordingly, the trajectory dominant zones are in either red (transit trajectory dominant) or white color (ridesharing trajectory dominant), even regions and unidentifiable pixels (see Definition 3) are in orange color with different intensity. In addition, Figure 9 (a) illustrates the solution of a 3D discretization, while Figure 9 (b) and (c) respectively present a heatmap (10:00:00 - 10:01:30) for the solution with $\kappa^* = 28^2$ and $\kappa^* = 16^2$ with the UI equal to 0.3 and 0.2. Note that there are too many cubes to be clearly outlined in Figure 10 (a), we instead show the height of cubes, τ^*, along the time dimension.

Table 2: Sensitivity analysis of UIs

UI	$(\underline{\eta}, \overline{\eta}, \underline{\mu}, \overline{\mu})$	κ^*	Total noise ratio
0.40	(0.05, 0.95, 0.45, 0.55)	34^2	0.62
0.30	**(0.10, 0.90, 0.40, 0.60)**	$\mathbf{28^2}$	**0.45**
0.20	(0.15, 0.85, 0.35, 0.65)	16^2	0.37
0.10	(0.20, 0.80, 0.30, 0.60)	10^2	0.13

We evaluate the merits of the heatmaps in terms of the land-use pattern and the ratio of noise pixels over all pixels in a 3D discretization solution. The noise pixels are not very interested in our analysis because they do not offer much insightful information for the service gaps for both service modes. The results in Table 2 indicate that the UI of 0.3 is better than the UI of 0.4 since the ratio of noise pixels is decreased from 0.62 to 0.45. This ratio can be further reduced from 0.45 to 0.13 as the UI decreases from 0.3 to 0.1 but with a significant loss of resolution. Specifically, we conduct the land-use analysis to examine the resolution of the heatmaps shown in Figure 9 (b) and (c). Region 1 and region 2 (marked out by the dashed outline) are chosen as the benchmark regions. The actual land-use analysis shows that region 1 is industrial land with a large waste disposal plant, several major intercity railway lines, and railway

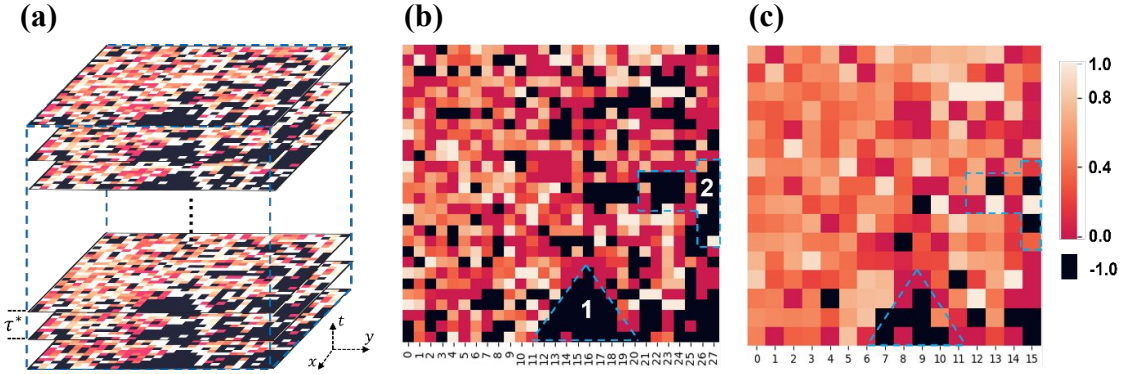

Figure 9. (a) Discretization of transit and ridesharing trajectory data in 3D space. Land-use analysis on heatmaps with different pixel numbers to evaluate resolution. (b) 10:00:00-10:01:30 heatmap with optimal pixel number 282 (0.3 UI). (c) 10:00:00-10:01:30 heatmap with optimal pixel number 162 (0.2 UI).

companies. Region 2 mainly consists of parks, intercity highways, and highway interchange junctions. These results indicate that both regions have few mobility needs and limited attractiveness for transit and ridesharing services. We next take a look at the heatmap results. The contours of these two regions are clearly outlined in Figure 9 (b) and they are mainly in black (i.e., no transport services provided). This is consistent with the actual land-use analysis. But, Figure 9 (c) does not demonstrate the same quality of the

resolution. Thus, we conclude that reducing UI from 0.2 to 0.1 significantly compromises the resolution of heatmaps. The above analysis confirms our best choice of UI (= 0.3) for this case study.

RS Zones and Insights

According to the approach developed in Section: *Searching RS zones*, we project the heatmaps $h_i, i \in I$ to a 2D spatial plane and then get the aggregated heatmap H. Built upon H, we apply the Algorithm 1 to recognize the ridesharing swarms (RS) zones on the heatmap H, with the threshold value $\bar{\eta}$=0.9 (the best value of $\bar{\eta}$ is evaluated from the sensitivity analysis in the previous section). We find six major RS zones. Combined with the POIs (points of interest) data, we overlay the RS zones to the POIs kernel density estimation (KDE) map in Figure 10 (a), where the RS zones are marked by the black line and the background intensity represents the density of commercial and residential POIs. The commercial and residential POIs mainly consist of residence communities, office buildings, and shopping centers. We observe that these RS zones mainly cover or near commercial or residential regions, except RS zone 2, which is a big theme park in Chengdu City. An existing study (Yu & Peng, 2019) has found that the areas with these types of land uses often generate high demand for on-demand services, and hybrid mobility service will be a good mobility solution. These results echo our conjecture that RS zones and the potential critical zones for hybrid mobility service are highly correlated.

Table 3: Number of trips per day between RS zones

RS zones pair	Mean/day	Standard Deviation	RS zones pair	Mean/day	Standard Deviation
(1,2)	250.9	17.6	(2,6)	5.0	1.7
(1,3)	**2531.4**	**243.5**	**(3,4)**	**2762.8**	**159.5**
(1,4)	463.7	76.2	**(3,5)**	**2878.5**	**396.2**
(1,5)	215.6	49.7	(3,6)	431.7	76.2
(1,6)	26.8	7.5	**(4,5)**	**7617.8**	**709.2**
(2,3)	172.0	15.1	(4,6)	131.0	12.7
(2,4)	655.9	94.8	**(5,6)**	**850.6**	**103.9**
(2,5)	333.6	40.8			

We further confirm this point by overlaying the RS zones to the transit stop density map as shown in Figure 10. It shows that the transit stop density is smaller in RS zones than in surrounding areas. We also investigated the number of ridesharing trips passing through these major RS zones (not necessarily covering the trips' origins and destinations) and show the results in Table 3. From the table, it is observed that a great number of daily ridesharing trips passing through RS zones 1 and 3 (2531.4 per day), RS zones 3 and 4 (2762.8 per day), RS zones 3 and 5 (2878.5 per day), and RS zones 4 and 5 (7617.8 per day). These results are consistent with the land use features of these RS zones. For example, zone 4 is a residential area near the major metro hub, thus it attracts significant and stable demand from zone 3 and zone 5. Given the existing high volume of ridesharing trips, these results indicate the pair of zone 4-3 suggests the candidate corridor zone, where the current transit service is insufficient (represented by the low transit density area between the zone pair 4-3 in Figure 10 (b)). For this corridor zone, we can consider implementing flexible transit routes/connections to improve the connectivity of the transit network (see the illustration in Figure 10 (b)). Moreover, the area between the pair of zone 4-5 has a high transit stop density, but has a connection gap to zone 4, which indicates the implementation of on-demand services is a potential solution for gap-closing, see Figure 10 (b). Note that our analysis provides the planning guidance, and we need more data to make the final operation decisions on these candidate solutions, which is out of the scope of this study.

On the other hand, our data analysis observes that the number of daily demands to zones 3 is around 36,000 per day, which counts for 19.5% of total ridesharing trips of the study area. Moreover, this zone is very large with low inside transit coverage and no walking-distance transit service connections. For example, zone 3 in Figure 10 (b) is about 2.69 mi^2. Given most of its pixels are in light color, we can say that zone 3 is of low transit trajectory density. Next, we can observe that the area adjacent to zone 3 (between zone 3 and 5) has a high transit stop density. Combining the features of zone 3 and these observations, we conclude that zone 3 lacks walking-distance transit services within the zone or to the surrounding zones, even though sufficient ad hoc demands exist. Therefore, they are good candidates to implement on-demand services, e.g., microtransit, ridesharing, or mini-bus services, for connecting to the nearby transit network.

Overall, the above results indicate that investigating the RS zones is very valuable. They help the planner narrow down the critical zones in a city for cooperating transit and on-demand services. Especially, by using the entire trajectory data, we can help find those candidate corridors that can be a potential flexible transit connection or microtransit route injecting to existing transit routes, but often cannot be discovered by only analyzing pick-up and drop-off data. Accordingly, these zones often do not get enough attention from the literature. The results also confirm the efficiency of our crowdsourced data analysis approach. Please note that this study focuses on helping narrow down the potential critical zones for the planner to consider implementing new services for connecting transit networks and on-demand services. To further judge which instrument fits best or how to implement them properly is out of the scope of this study.

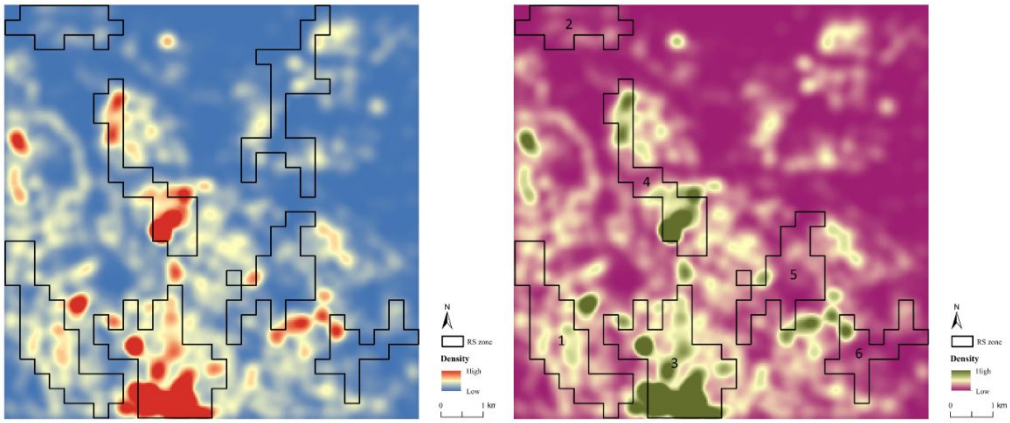

Figure 10. (a) Identified RS zones on POIs density map. (b) Identified RS zones on bus stop density map.

Inferring FLM Zones

We next validate the correlation between the "sandwich pattern" and FLM zones. To do that, we average every 40 (which is flexible, selected according to the data) heatmaps h_i, $i \in I$, each in 90 seconds, to generate the aggregated heatmap $\hat{h}_\omega$, $\omega \in \mathcal{W}$. Therefore, each heatmap $\hat{h}_\omega, \omega \in \mathcal{W}$ contains the trajectory information in one hour. Upon $\hat{\boldsymbol{h}} = \{\hat{h}_\omega, \omega \in \mathcal{W}\}$, we apply Algorithm 2, with $\overline{\eta} = 0.9$ and $\underline{\eta} = 0.1$ (values are selected from the sensitivity analysis in Section: *Establishing 3D discretization*) to recognize the "sandwich" patterns and output the FLM-prone order probability heatmaps $\tilde{h}_\omega \in \tilde{\boldsymbol{h}}$ by integrating the ridesharing O-D information and transit stop data. Figure 11 (a) presents an example of the FLM probability heatmap (9 AM to 10 AM), in which some FLM zones are marked by the yellow dash cycle. The pixel with lighter color possesses more FLM-prone orders. Since the FLM demand is limited, Figure 11 (a) is almost

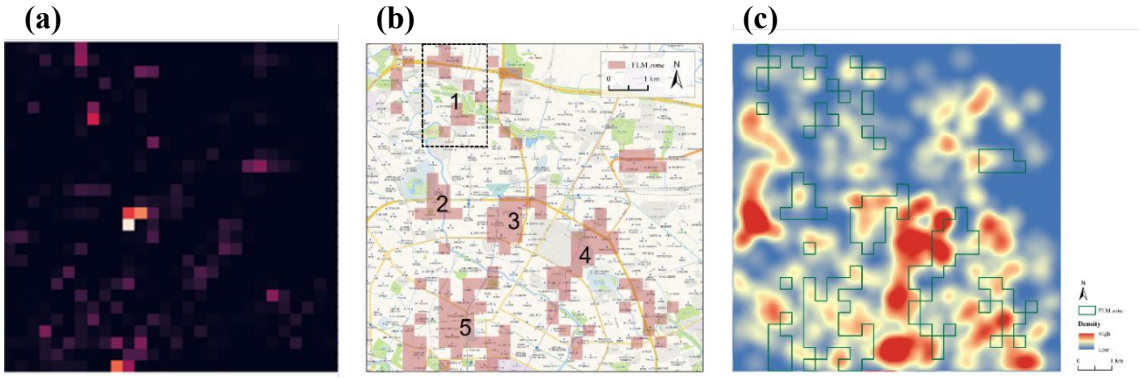

Figure 11. (a) FLM-prone order probability heatmap; (b) Identified FLM zones on map; (c) Overlap of FLM zones and transit stops distribution heatmap.

black. For better visualization, Figure 11 (b) presents the major FLM zones identified on the map.

We then examine the FLM zones founded by the "sandwich pattern". Due to the lack of intermodal trip data, we validate these FLM zones by using the two features to characterize an FLM zone in existing

literature, (i) having mixed land-use with sufficient travel demands; (ii) low transit service coverage.

With this consideration, we first compare the transit service density within the FLM zones to the surrounding areas by mapping the FLM zones to the transit stop density map, as shown in Figure 11 (c). It is observed that the identified FLM zones have higher FLM-prone orders as shown in Figure 11 (a), but have lower transit service density than the surrounding areas as indicated by Figure 11 (c). This observation is consistent with the features claimed in the literature and validates the FLM zones we found (Guo & He, 2020; Mo et al., 2018).

Next, we consider that a promising FLM zone should attract enough demand. This motivates us to validate the FLM zones by examining the land-use of the study region. Accordingly, we investigate the relationship between FLM-prone order probability and the percentage/number of commercial and residential POIs. The results are presented in Figure 12, in which we can observe that the number of commercial and residential POIs increases with the FLM-order probability. It indicates that the FLM zones have more commercial and residential POIs than other zones and it usually suggests a higher population and FLM demands. Moreover, compared with the Open Street map in Figure 11 (b), we can see that zones 1 – 5 are located around metro lines. These land-use features reinforce our analysis that they are most likely FLM zones not well connected to the existing backbone transit lines. Therefore, the on-demand mobility service such as ridesharing or microtransit is a good complement to the backbone transit lines. Except zones 1 – 5, other FLM zones are scattered within suburban areas, with low transit services and are far away from a metro line. They are typical transit deficiency zones, where new transit routes are needed to make up the service gaps. Overall, the validation shows that the FLM zones we discovered by the "Sandwich" pattern are consistent with the features claimed in the literature (Guo & He, 2020; Mo et al., 2018). Therefore, the case study demonstrates that our approach works efficiently to find potential FLM zones and provides valuable guidance to fulfil the service gaps.

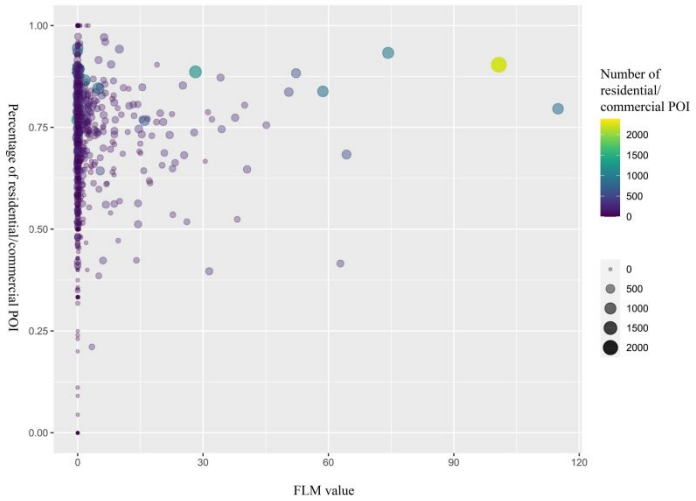

Figure 12. Land use validation for FLM zones. Percentage/number of commercial and residential POIs versus FLM-prone order probability.

Predicting FLM

We next demonstrate the performance of the machine learning model to predict the FLM zones. Specifically, the input data is a two-channel tensor and we predict the dynamics of RS zones through the first channel and the spatial change of FLM locations over time by the second channel. To do that, we implement the ConvLSTM using the framework proposed by (Xingjian et al., 2015), using the Keras API. It consists of five layers. The first and third layers are ConvLSTM with 120 size 2 × 2 kernels and 80 size 3 × 3 kernels, respectively. The second and fourth layers are batch normalization. Finally, we used 2 × 2 × 2 kernels to get the output shape. We use Mean Average Percentage Error (MAPE) and Location Prediction Error (LPE) to evaluate the performance of the ConvLSTM network, which are defined as follows:

$$MAPE = \frac{1}{\kappa}\sum_{k=1}^{\kappa}\frac{|\hat{y}_{i+1}^{k} - y_{i+1}^{k}|}{y_{i+1}^{k}} \quad (11)$$

$$LPE = \frac{1}{\kappa}\sum_{k=1}^{\kappa}\left|\hat{\delta}_{i+1}^{k} - \delta_{i+1}^{k}\right| \quad (12)$$

where $\hat{y}_{i+1}^{k}$, y_{i+1}^{k} are the prediction and the real value of pixel k for time interval $i+1$. $\hat{\delta}_{i+1}^{k}$ and δ_{i+1}^{k} take 0-1 values and respectively indicates whether pixel k is an FLM zone in the prediction or field data. κ is the total number of pixels in an image. The goal of the FLM prediction is to identify where is the potential FLM zones in the near future time interval. Therefore, LPE is adopted to evaluate the performance of the FLM prediction. A smaller value of the LPE indicates a higher prediction accuracy.

The training data consists of the heatmap $\hat{h}_{\omega} \in \hat{\boldsymbol{h}}$ as the first channel input and the corresponding heatmap $\tilde{h}_{\omega} \in \tilde{\boldsymbol{h}}$ as the second channel input at each time interval $\omega \in \mathcal{W}$. Each time interval $\omega \in \mathcal{W}$ has one-hour interval width. The output is the two-channel prediction: $\hat{h}_{\omega+1}$ and $\tilde{h}_{\omega+1}$ in the next time interval (next hour). The model is trained with the first three weeks of data. The data of last week is used for validation. The model achieves 28.42% (MAPE) for the first channel and 4.22% (LPE) for the second channel. Therefore, the model exhibits high accuracy in predicting the locations of FLM zones at future intervals, which provides a promising radar for the future FLM demand and the corresponding needs for ridesharing or microtransit services.

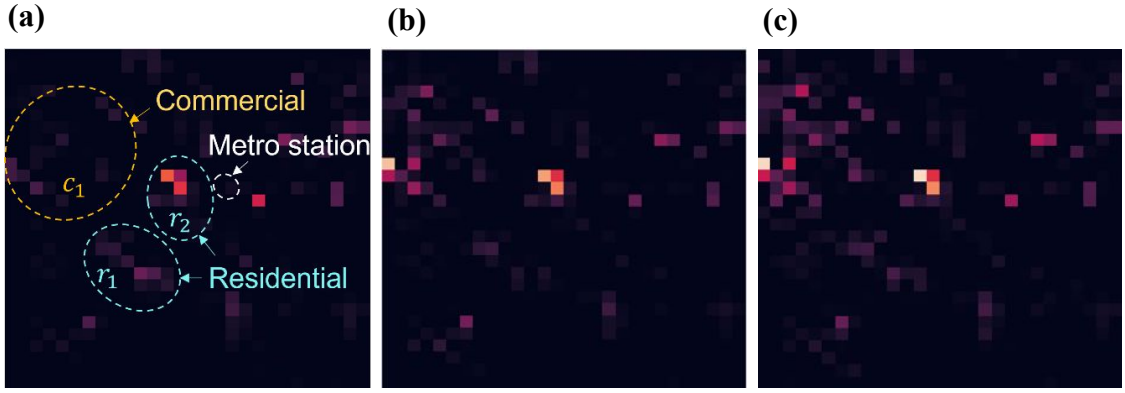

Figure 13. Prediction of time-vary FLM heatmap. FLM heatmap of (a) 8:00-9:00, (b) 9:00-10:00, (c) 18:00-19:00. Lighter pixel indicates a higher possibility of FLM demand.

Moreover, the prediction results also demonstrate the time-variant characteristics of the FLM demand. For example, Figure 13 (a) to (c) indicates the interesting dynamics of the FLM demand during the period covering the morning peak hour (8:00-9:00), secondary-peak hour (9:00-10:00), and evening peak hour (18:00-19:00). More exactly, it is noticed that some zones such as c_1 on the upper-left corner presents lower FLM demand during the morning peak hour, while other zones, such as r_1, have more FLM demand during the morning and evening peak hours than that in the secondary-peak hour. Moreover, some zones, such as r_2, consistently have high FLM demand. We further validate the rationality of these predicted dynamics by examining the land-use patterns. We found that zone c_1 involves a group of big shopping centers. Given most of the shopping malls are closed from 8:00 to 9:00, it is reasonable to have less FLM demand in this time range. Both r_1 and r_2 are residential zones, but r_2 has a large metro hub nearby. The predicted temporal FLM demand pattern of r_1 is consistent with the demand features that fewer work trips occur during non-peak hours compared with the peak hours. And the metro hub near r_2 constantly generates significant FLM demand, which is also consistent with the prediction. The stable FLM demand in r_2 calls for the improvement of transit coverage by either creating a new transit stop (inflexible service) or providing microtransit (flexible service) as feeders to the transit backbone nearby. In conclusion, our prediction is consistent with the land-use analysis, which confirms the effectiveness of the machine-learning model to predict the potential FLM demand.

Conclusion

Considering the united and large-scale transit and ridesharing trajectory data collected by many agencies as the crowdsourced data, this study developed an innovative data analysis approach to discover the critical transfer zones, where cooperating transit and on-demand services will potentially improve the mobility services. To discover this knowledge hidden in the well-collected trajectory data, we first meshed the trajectory data into an optimal 3D discretization with a uniform cube size. Then we zoomed in and

investigated the mobility service information in cubes in each time interval to form the heatmaps. Built upon the heatmaps, we zoomed out and discovered two important patterns, which exhibit great values to infer the critical zones for cooperating transit and on-demand services. Mainly, we examined the ridesharing swarm zones, which show a strong correlation with the promising hotspots and corridor zones needing such hybrid mobility services. Next, we investigated the "sandwich" patterns on the aggregated heatmaps to discover potential FLM zones. Further, by feeding the heatmaps into a two-channel ConvLSTM model, this analysis predicted the time-varying FLM demands and identified potential FLM zones. A case study conducted for the second ring region of Chengdu, China validated the effectiveness and capability of our analysis approach. The FLM demand prediction helps adapt the public transport system to the time-varying ad-hoc FLM demands by implementing hybrid mobility services such as microtransit. For example, it can serve as "demand radar" for microtransit schedules planning. Moreover, our data analysis approach helps find the candidate critical zones for implementing hybrid mobility service or transit network redesign in a large urban area. It benefits the transit agencies because our approach can serve as the first step to narrow down the planning zones so that the limited planning resources can be better exploited. We will explore potential future research that stems from this study in several directions. First of all, the crowdsourced data analysis approach can be extended by involving the trajectory data collected from other mobility services correlated to transit systems such as micro-mobility services and private auto. It will help discover the critical zones by fully investigating the existing mobility supply. Moreover, it is very interesting to leverage the critical zones discovered by this study in the transit operation planning and network redesign problems. Also, with our optimal discrete representation approach, we are able to understand the mobility patterns in a more spatially and temporally resolute way, which helps us to develop more effective and flexible strategies in supply management and better serve the dynamic travel demand. This future work will optimally determine what and how on-demand mobility services should be adopted at each candidate zone with the aim to maximize the system benefit. We believe the approach developed in this study can help shrink the solution searching space for the corresponding decision models. Last, we believe our analysis method can be transferred to other planning problems, such as bikesharing station planning, to account for and benefit from the integration of multiple transportation modes.

References

Aldaihani, M. M., Quadrifoglio, L., Dessouky, M. M., & Hall, R. (2004). Network design for a grid hybrid transit service. *Transportation Research Part A: Policy and Practice*, *38*(7), 511–530.

Bastani, F., Huang, Y., Xie, X., & Powell, J. W. (2011). A greener transportation mode: flexible routes discovery from GPS trajectory data. *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 405–408.

Berlingerio, M., Calabrese, F., di Lorenzo, G., Nair, R., Pinelli, F., & Sbodio, M. L. (2013). AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 663–666.

Berrada, J., & Poulhès, A. (2021). Economic and socioeconomic assessment of replacing conventional public transit with demand responsive transit services in

low-to-medium density areas. *Transportation Research Part A: Policy and Practice*, *150*, 317–334.

Boarnet, M. G., Giuliano, G., Hou, Y., & Shin, E. J. (2017). First/last mile transit access as an equity planning issue. *Transportation Research Part A: Policy and Practice*, *103*, 296–310.

Boyle, D. K. (2006). *Fixed-route transit ridership forecasting and service planning methods* (Vol. 66). Transportation Research Board.

Cao, L., & Krumm, J. (2009). From GPS traces to a routable road map. *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 3–12.

Castro, P. S., Zhang, D., & Li, S. (2012). Urban traffic modelling and prediction using large scale taxi GPS traces. *International Conference on Pervasive Computing*, 57–72.

Chatterjee, A., & Venigalla, M. M. (2004). Travel demand forecasting for urban transportation planning. *Handbook of Transportation Engineering*, *1*.

Chen, C., Zhang, D., Zhou, Z.-H., Li, N., Atmaca, T., & Li, S. (2013). B-Planner: Night bus route planning using large-scale taxi GPS traces. *2013 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 225–233.

Dechter, R., & Pearl, J. (1985). Generalized best-first search strategies and the optimality of A. *Journal of the ACM (JACM)*, *32*(3), 505–536.

Faghih, S. S., Safikhani, A., Moghimi, B., & Kamga, C. (2019). Predicting Short-Term Uber Demand in New York City Using Spatiotemporal Modeling. *Journal of Computing in Civil Engineering*, *33*(3), 5019002.

Fang, Z., Cheng, Q., Jia, R., & Liu, Z. (2018). Urban rail transit demand analysis and prediction: A review of recent studies. *International Conference on Intelligent Interactive Multimedia Systems and Services*, 300–309.

Fu, L. (2002). Planning and design of flex-route transit services. *Transportation Research Record*, *1791*(1), 59–66.

Grahn, R., Qian, S., & Hendrickson, C. (2022). Optimizing first-and last-mile public transit services leveraging transportation network companies (TNC). *Transportation*, 1–28.

Guo, Y., & He, S. Y. (2020). Built environment effects on the integration of dockless bike-sharing and the metro. *Transportation Research Part D: Transport and Environment*, *83*, 102335.

Hadjidimitriou, N. S., Lippi, M., & Mamei, M. (2020). A Data Driven Approach to Match Demand and Supply for Public Transport Planning. *IEEE Transactions on Intelligent Transportation Systems*.

Hashemian, H. (2002). Using GIS to Assess Demographic and Land Use Characteristics on Local Transit Services. *Seventh TRB Conference on the Application of Transportation Planning MethodsTransportation Research Board; Commonwealth of Massachusetts, Executive Office of Transportation and Construction; and Boston Metropolitan Planning Organization.*

Huang, H. (1996). The land-use impacts of urban rail transit systems. *Journal of Planning Literature*, *11*(1), 17–30.

Johnson, S., Zalewski, A., Eby, B., & Lewis, P. (2020). *Redesigning Transit Networks for the New Mobility Future : Resource and Toolkits DRAFT Final Report*. https://doi.org/10.17226/26028

Jun, M.-J., Choi, K., Jeong, J.-E., Kwon, K.-H., & Kim, H.-J. (2015). Land use characteristics of subway catchment areas and their influence on subway ridership in Seoul. *Journal of Transport Geography*, *48*, 30–40.

Koffman, D. (2004). Operational experiences with flexible transit services. *Transportation Research Board of the National Academies*, *TCRP Synth*.

Li, J.-Q., Zhou, K., Zhang, L., & Zhang, W.-B. (2012). A multimodal trip planning system with real-time traffic and transit information. *Journal of Intelligent Transportation Systems*, *16*(2), 60–69.

Li, X., & Quadrifoglio, L. (2010). Feeder transit services: choosing between fixed and demand responsive policy. *Transportation Research Part C: Emerging Technologies*, *18*(5), 770–780.

Liu, L., Qiu, Z., Li, G., Wang, Q., Ouyang, W., & Lin, L. (2019). Contextualized spatial--temporal network for taxi origin-destination demand prediction. *IEEE Transactions on Intelligent Transportation Systems*, *20*(10), 3875–3887.

Luo, S., & Nie, Y. M. (2019). Impact of ride-pooling on the nature of transit network design. *Transportation Research Part B: Methodological*, *129*, 175–192.

Maheo, A., Kilby, P., & van Hentenryck, P. (2019). Benders decomposition for the design of a hub and shuttle public transit system. *Transportation Science*, *53*(1), 77–88.

Mo, B., Shen, Y., & Zhao, J. (2018). Impact of built environment on first-and last-mile travel mode choice. *Transportation Research Record*, *2672*(6), 40–51.

Mounce, R., Wright, S., Emele, C. D., Zeng, C., & Nelson, J. D. (2018). A tool to aid redesign of flexible transport services to increase efficiency in rural transport service provision. *Journal of Intelligent Transportation Systems*, *22*(2), 175–185.

Nazem, M., Trépanier, M., & Morency, C. (2011). Demographic analysis of route choice for public transit. *Transportation Research Record*, *2217*(1), 71–78.

Noursalehi, P., Koutsopoulos, H. N., & Zhao, J. (2018). Real time transit demand prediction capturing station interactions and impact of special events. *Transportation Research Part C: Emerging Technologies*, *97*, 277–300.

Potts, J. F., Marshall, M. A., Crockett, E. C., & Washington, J. (2010). *A guide for planning and operating flexible public transportation services*.

Qiu, F., Li, W., & Zhang, J. (2014). A dynamic station strategy to improve the performance of flex-route transit services. *Transportation Research Part C: Emerging Technologies*, *48*, 229–240.

Quadrifoglio, L., & Li, X. (2009). A methodology to derive the critical demand density for designing and operating feeder transit services. *Transportation Research Part B: Methodological*, *43*(10), 922–935.

Rahimi, M., & Dessouky, M. (2001). A hierarchical task model for dispatching in computer-assisted demand-responsive paratransit operation. *Journal of Intelligent Transportation Systems*, *6*(3), 199–223.

Roberts, R. A. (1985). Analysis of demographic trends and travel patterns: Implications for the future of the Portland transit market. *Transportation Research Record: Journal of the Transportation Board*, *1067*, 1–8.

Shen, Y., Zhang, H., & Zhao, J. (2018). Integrating shared autonomous vehicle in public transportation system: A supply-side simulation of the first-mile service in Singapore. *Transportation Research Part A: Policy and Practice*, *113*, 125–136.

Shu, P., Sun, Y., Xie, B., Xu, S. X., & Xu, G. (2021). Data-driven shuttle service design for sustainable last mile transportation. *Advanced Engineering Informatics*, *49*, 101344.

Stiglic, M., Agatz, N., Savelsbergh, M., & Gradisar, M. (2018). Enhancing urban mobility: Integrating ride-sharing and public transit. *Computers & Operations Research*, *90*, 12–21.

Sung, H., Choi, K., Lee, S., & Cheon, S. (2014). Exploring the impacts of land use by service coverage and station-level accessibility on rail transit ridership. *Journal of Transport Geography*, *36*, 134–140.

Teal, R. F. (1994). Using smart technologies to revitalize demand responsive transit. *Journal of Intelligent Transportation System*, *1*(3), 275–293.

Trépanier, M., Tranchant, N., & Chapleau, R. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, *11*(1), 1–14.

Velaga, N. R., Nelson, J. D., Wright, S. D., & Farrington, J. H. (2012). The potential role of flexible transport services in enhancing rural public transport provision. *Journal of Public Transportation*, *15*(1), 7.

Wang, R., Chen, F., Liu, X., Liu, X., Li, Z., & Zhu, Y. (2021). A Matching Model for Door-to-Door Multimodal Transit by Integrating Taxi-Sharing and Subways. *ISPRS International Journal of Geo-Information*, *10*(7), 469.

Xingjian, S. H. I., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, 802–810.

Xu, J., Rahmatizadeh, R., Bölöni, L., & Turgut, D. (2017). Real-time prediction of taxi demand using recurrent neural networks. *IEEE Transactions on Intelligent Transportation Systems*, *19*(8), 2572–2581.

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. *Journal of Transport Geography*, *75*, 147–163.

Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G., & Huang, Y. (2010). T-drive: driving directions based on taxi trajectories. *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 99–108.

Zhang, J., Zheng, Y., & Qi, D. (2016). Deep spatio-temporal residual networks for citywide crowd flows prediction. *ArXiv Preprint ArXiv:1610.00081*.

Zhang, K., Liu, Z., & Zheng, L. (2019). Short-term prediction of passenger demand in multi-zone level: Temporal convolutional neural network with multi-task learning. *IEEE Transactions on Intelligent Transportation Systems*, *21*(4), 1480–1490.

Zheng, Y., Liu, Y., Yuan, J., & Xie, X. (2011). Urban computing with taxicabs. *Proceedings of the 13th International Conference on Ubiquitous Computing*, 89–98.

Zhou, X., Shen, Y., Zhu, Y., & Huang, L. (2018). Predicting multi-step citywide passenger demands using attention-based neural networks. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 736–744.

Appendix

FLM zones search algorithm.

Algorithm 2 FLM Zones Search

Procedure *FLMZONES*($\hat{h}_\omega$, $\overline{\rho}$)

Create the FLM probability heatmap $\tilde{h}_\omega$ with the heat $\tilde{r}_{k,\omega} = 0, \forall k \in K$

Generate trajectory density heatmap

$\mathcal{A} \leftarrow$ *CLUSTER*($\hat{h}_\omega$, $\overline{\eta}$) # clustering criterion: $\hat{r}_{k,\omega} \geq \overline{\eta}$

$\mathcal{B} \leftarrow$ *CLUSTER*($\hat{h}_\omega$,$\underline{\eta}$) # clustering criterion: $\hat{r}_{k,\omega} \leq \underline{\eta}$

Identify "sandwich" patterns

for $A_k \in \mathcal{A}$ **do**

  if $\rho(A_k) > \overline{\rho}$ **then**

    $\mathcal{A} \leftarrow \mathcal{A} \setminus \{A_k\}$

    for $B_k \in \mathcal{B}$ **do**

      if $\rho(B_k) < \overline{\rho}$ **then**

        $\mathcal{B} \leftarrow \mathcal{B} \setminus \{B_k\}$

```
    for B_k ∈ ℬ do
        B_k^A ← ∅
        for ĥ_ω(l) ∈ B_k do
          for A_k ∈ 𝒜 do
            if O_l ∩ A_k ≠ ∅ then
              B_k^A ← B_k^A ∪ A_k
  # Find FLM-prone orders
    for B_k ∈ ℬ do
      for A_k ∈ B_k^A do
        for ĥ_ω(m) ∈ A_k do
          for A_l ∈ B_k^A do
            for ĥ_ω(l) ∈ A_l do
              for v ∈ V_ω do
                if o_v ∈ ĥ_ω(m) and d_v ∈ ĥ_ω(l) and A_k ≠ A_l then
                  γ_m += 1 and γ_l += 1
    for B_k ∈ ℬ do
      for A_k ∈ B_k^A do
        for ĥ_ω(l) ∈ A_k do
          for v ∈ V_ω do
            if o_v ∈ A_k and d_v ∈ O_l ∩ B_k then
            γ_l += 1
        for ĥ_ω(m) ∈ A_k do
          for ĥ_ω(l) ∈ B_k do
            for v ∈ V_ω do
              if o_v ∈ ĥ_ω(l) and d_v ∈ ĥ_ω(m) then
                γ_m += 1
  # Create FLM-prone order probability heatmap
  for k ∈ K do
    r̃_{k,ω} = γ_k / Σ_k γ_k
  return h̃_ω
```