Mobile Phone Data in Urban Customized Bus: A Network-based Hierarchical Location Selection Method with an Application to System Layout Design in the Urban Agglomeration

Qing Yu 1,2, Weifeng Li 1,*, Haoran Zhang 2 and Dongyuan Yang 1

1 Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, 4800 Cao’an Road, Shanghai 201804, China; qingyu@csis.u-tokyo.ac.jp (Q.Y.); yangdyk@tongji.edu.cn (D.Y.)
2 Center for Spatial Information Science, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba 277-8568, Japan; zhang_ronan@csis.u-tokyo.ac.jp
* Correspondence: lwf0103@gmail.com

Received: 28 June 2020; Accepted: 30 July 2020; Published: 31 July 2020

Abstract: Employment is one of the essential socioeconomic connections in urban agglomeration. However, both the demand and the supply of transjurisdictional public transport service are unevenly distributed in such area. Customized bus has a high potential of serving transjurisdictional and long-distance commuting demand. This study proposes a network-based layout design method to generate hierarchical service scopes and stations for customized bus system. First, the home and the workplace of residents are identified using mobile phone data, to construct a jobs-housing relationship network. Then, an iterative algorithm based on network community detection and density-based spatial clustering is applied to the jobs-housing relationship network to hierarchically segment the urban agglomeration area into communities. Three methods are proposed for location selection of customized bus stations. A case study is conducted using the mobile phone dataset from nine cities in the Yangtze River Urban Agglomeration in China. A four-level hierarchical customized bus system layout is generated and both the spatial properties and network properties of service scopes are analyzed. The proposed three methods of customized bus station location selection are compared based on the average travel distance and the rationality of the resulted customized bus station location.

Keywords: customized bus; system layout design; hierarchical continuous location problem; urban agglomeration transportation

1. Introduction

Urban land expansion and urban population growth are the outcomes of rapid urbanization. In China, the urbanization rate increased from 17.4% to 58.5% between 1978 and 2017, and it is expected to reach 75% by 2050. With the massive population flooding into the urban areas, urban agglomeration is widely believed as the future direction of further urbanization and regional development. Urban agglomeration is a highly developed spatial form of integrated cities. The occurrence of urban agglomeration means that the relationships among cities shift to both competition and cooperation. Unlike the morphological patterns of individual cities, urban agglomeration includes not only the urban area but also satellite cities and intervening rural land [1,2]. These outlying zones, not necessarily urban in character, are closely bound to the core urban
areas by employment or other commerce. The resulted jobs-housing relationship is known as a commuter belt and may extend well beyond the urban zone, to other political entities.

However, the development of the transportation system in urban agglomeration is extremely uneven. In terms of supply, for central urban areas in the urban agglomeration, public transportation networks are usually well developed and connected, offering high accessibility for citizens. On the other hand, in the peripheral residential areas, the development of the public transportation system is usually lagging [3]. The poor transport connections at the home end, as well as the absence of efficient public transportation to potential destinations, can prevent people from transjurisdictional traveling and lead to social exclusion. In terms of demand, according to the commuting demand revealed by the mobile signaling data collected in the Yangtze River Delta, there are 57,000 transjurisdictional commuters living or working in Shanghai in 2019[4]. The unevenly distributed demand for commuting makes it difficult to offer direct and efficient public transport services in the urban agglomeration, especially encountered with the jurisdiction under different political entities and the conflict of interest between different entities.

Thus, it is urgent to offer a fast, direct, and efficient transportation mode for commuting as a supplement of the unevenly developed transportation system in urban agglomeration. Commuting oriented customized bus (CB) can be a perfect solution as a supplement of the unevenly developed transportation system in the urban agglomeration. CB offers opportunities to serve the sharing demand for groups of passengers with similar travel demand. Comparing with traditional bus lines, a CB system requires passengers to submit their origin, destination and departure time via the internet to reserve seats. The design of CB lines emphasizes the directness with few or no intermediate stops along the route. Dozens of papers have also shown that commuting trips have high potential and acceptance of sharing transportation mode [5,6]. The high reliability, directive, and efficiency of CB lines make it a perfect solution for long-distance commuting demand in the urban agglomeration area. A well-designed CB system that matches the actual commuting demand will be profitable and readily accepted by the public.

This paper introduces a CB system layout design and develops a network-based hierarchical location selection method based on the jobs-housing relationship from mobile phone data. The rest part of this paper is organized as follows—hereafter this section, this paper reviews the related works from the literature and describes the overview of the CB system and the basic rules of the service scope division and station location selection; Section 2 presents the detailed methodology of this work, including the iterative algorithm of generating service scopes and station location; Section 3 presents the result of the case study in Yangtze River Urban Agglomeration in China; Section 4 proposes the conclusions and the future prospects of this study.

1.1. Related Works

A well-organized public bus service can attract more passengers and reduce the explicit dependency on private cars [7]. As a type of public transportation mode, the concept of CB service was initially introduced as subscription bus services several decades ago. Kirby and Bhatt introduced the concept of subscription bus services and developed guidelines for subscription bus networks, including operational planning and pricing strategy [8]. Bautz described several types of subscription bus and recommended to improve the climate for the growth of subscription service [9]. McCall analyzed the development and operation of a subscription bus system, which had been successfully implemented in Ventura, Los Angeles, and Orange Village, California [10].

Recently, Internet-supported customized bus mode had emerged. Qingdao launched the first CB system in August 2013 in China and gradually became warmly applauded by the public. Liu et al. introduced the concept of the new customized bus and provided a systematic examination and analysis of the development and current state of CB practices in China [11].

The design of the newly developed CB system also attracted the attention in recent studies. These studies mainly focused on the basic design of CB system [12], including bus line routing and station location design, bus timetable scheduling, crew and vehicle dispatching, and bus real-time control.
System layout design is a crucial stage to optimize a system and a strategy to improve system performance [13,14]. Optimization models also have been proposed to optimize the layout design of CB systems. These models usually follow several specific steps as: 1. demand extraction, 2. CB line and stops deployment, 3. further optimization on scheduling, dispatching and controlling, etc. In the detailed design of CB systems, studies well-developed dozens of models that involve multiple factors, including dynamic routing and timetabling, congestion detouring, etc. Ma et al. proposed an immune genetic algorithm-based model to determine CB stations and CB timetables [15]. Yu et al. introduced a graph-based method to generate planning suggestions for CB line and stops based on massive mobile phone trajectory data [16]. Tong et al. developed a joint optimization model for the planning and optimizing of detailed bus routing and timetabling plans [17]. Aiming to integrate CB system into the intelligent transportation system, Chuanyu et al. proposed a set of intelligent dispatching strategies for CB to mitigate traffic congestions [18]. Lyu et al. developed a bus line planning framework for CB systems, which simultaneously optimized bus stop locations, bus routes, timetables, and passengers’ probabilities of choosing CB [19].

Most of the existing studies with complicated optimization models are computationally expensive and mainly focus on the CB system design in a relatively small region. Applying these models to the design of the CB system in the metropolitan area or urban agglomeration area remains significant challenges:

1. First, the actual travel demand failed to be considered. In most existing CB systems, CB lines are planned based on travel demand data collected from on-line surveys, and the CB system layouts are designed based on the administrative division and organized by single cities. In the urban agglomeration area, survey data can hardly cover the area on a large scale. Moreover, even more importantly, the long tail travel demand in the urban agglomeration area, such as the long-distance commuting demands is unlikely to be discovered by the traditional survey data.

2. Second, the hierarchical character of travel demand failed to be considered. The transportation demands in the urban agglomeration area are comprehensive [20], consisting of trips in cross-city, intrametropolitan, and intracity levels. The traveling patterns of residents are much complicated and generate more frequent cross-city movement. Without considering the entirety of urban agglomeration and hierarchically design transportation systems for travel demand, it is unable to build an efficient system to serve the actual demand.

Seldom of the studies consider the overall layout design of the CB system. Nevertheless, in the urban agglomeration, the design of the CB system layout plays a crucial role in simplifying the problem before implementing CB line planning. In fact, in the currently operating CB service, the system layout design is commonly considered. For example, e-drive in Shanghai offers several types of CB services including long-distance CB service called microrailway, park area CB service, personal CB service, etc. In their microrailway service, e-drive divides the area of Shanghai into 11 subareas, offering 45 CB lines to connect between subareas [21]. Spatial area division is one of the methods to consider the demand [22,23]. Ma et al. emphasize that spatial area division is a component of the CB line and is an essential part of network planning [22]. In the urban agglomeration, a hierarchical CB system is a prerequisite for establishing a CB network that well fits the demand.

Passenger preference on CB system is also an important research aspect. Cao and Wang studied the passenger preference on CB using the SP (Stated Preference) and RP (Revealed Preference) survey to measure the individual willingness to choose CB [24]. Liu et al. assessed and compared the performance of CB with private cars and conventional public transport systems, respectively [25]. By assessing the overall performance of CB and comparing it with other types of transportation modes, Zhang et al. analyzed the provision of CB by applying analytical approaches based on demand analysis and network modeling and dedicated to find the key factors that significantly affect passenger mode shift performance [26].

The emerging pervasive, geospatial data generated by individuals has recently triggered an opportunity for studying individual mobility patterns [27]. The newly arisen datasets were widely used to obtain and evaluate the demand for new transportation modes, such as mobile phone data
[14,28,29], taxi GPS data [30], bus smart card data [31,32], etc. As for the recent studies on CB planning, this kind of human mobility data is also widely used since it can precisely capture the travel information at a low cost. Lyu et al. developed a CB line network planning framework called T2CBS which can discover similar travels from taxi trajectory data [33]. Guo et al. proposed a methodology to extract potential travel demand of CB using the bus smart card data [34].

1.2. Problem Description

1.2.1. Overview of the Dynamic CB Planning System

To offer a CB service, it is necessary to develop a system with dynamic CB planning. We provide an overview of the system in Figure 1. As the figure shows, first, passengers submit the traveling demand to the CB system. In the following steps, CB system gathers the massive demand data as the input of the system layout design and generates station location selection. For operation efficiency, the stations will be classified into a hierarchical structure and connected. Finally, the CB system will publish CB riding information to passengers and inform them of the location and time to get on the bus. In this planning system, the CB lines can be dynamic without fixed routes.

![Figure 1. Overview of a dynamic customized bus planning system.](image)

In this paper, the service scope of the CB system is defined as the acceptable range of areas for travelers to choose the CB as a transportation mode to complete their travel purposes. In the dynamic CB planning, the planning of CB system layout includes the determination of service scopes and the selection of station location. The accessibility and availability are two essential principles in the CB system planning. The accessibility of CB service is mainly embodied in the distance between the CB stations and the actual demand. The availability of CB service means that the determined service scopes should cover most of the actual demand. However, this problem is challenging to solve. It involves coupling problems of hierarchical demand clustering and continuous Multi-facility Weber Problem (MFWP) on massive data. The MFWP deals with finding the location of m facilities in continuous space, which has been proved as an NP-hard problem [35,36]. The objective function of MFWP is neither concave nor convex and may contain multiple local minima [37]. In the customized bus scenario, the method to design the system layout should be dynamic due to the constant change of the demand, which requires the high computing speed of the method on massive data.

Here, a heuristic algorithm for this problem is provided. A CB system layout with hierarchical service scopes is designed (see Figure 2). Commuting demand in the urban agglomeration is a combination of demand from multilevels. In order to serve the demand from different levels, as shown in the figure, the commuting CB system is designed to have multiple levels, including intercity connection (level 1), intracity connection (level 2), intracommunity connection (level 3), etc. In each level, the city regions are split into densely connected communities as the service scope. Each service scope possesses a single CB station that CB lines can be deployed to connect the service scopes. Taken each service scope into consideration, it can be divided into multiple smaller and more densely connected communities as subservice scopes in the sublevel. Moreover, CB stations and CB lines in the sublevel can be deployed and connected using the same method. In each level, the CB system mainly serves the transportation demand between different scopes at the same level, and the transportation demand inside each service scope is served by the sublevel CB service.
Figure 2. Overview of the hierarchical customized bus (CB) system.

The advantages of designing such a hierarchical system are as follows. 1. The problem is simplified into a hierarchical service scopes division problem and easy to be solved. Sublevels can be continuously segmented into smaller service scopes. With the sublevels continuously generated, the area connected by stronger commuting demand will generate more subservice scopes and sublevel stations. 2. The transportation connection served by the CB system will be a structured network with different density according to the real-world demand. Commuting demand at a higher level is less than the lower level. For example, the demand for intercity commuting is much less than that of intracity commuting. Correspondingly, in the hierarchical CB system, the number of service scopes at a higher level is much less than that at lower levels, which makes a denser CB network in the lower levels.

1.2.2. Customized Bus Service Scope Division

In each level of the CB system, the service scope is divided into subservice scopes. Here, two basic rules for service scope of commuting-oriented CB services are defined:

A service scope should be an area with a strong intraconnection of jobs-housing relationship. The connection between service scopes should be significantly less than that inside the service scope. For the great demand in the service scopes at a higher level, the hierarchical system can offer denser connection in sublevels; for small service scopes at the lowest level, the area is too small that it no longer needs the CB system.

A service scope should be a spatial cluster of regions. In real-world situations, a group of regions may be geospatially discrete but still generate a dense jobs-housing connection. These groups of regions will not be suitable to become a CB service scope.

Under these rules, a service scope will be a geospatially continuous region with a dense jobs-housing connection. Otherwise, the CB system can hardly assemble passengers to achieve desired operational performance.

1.2.3. Customized Bus Station Location Selection

Given a service scope segmentation in the CB system, the methodology aims to find a location to deploy bus stations in each service scope. Specifically, for each service scope, the following inputs are considered:
A set of the home location of cross-region commuters $H = \{hloc_i(x_i, y_i)\}$, in which $hloc_i(x_i, y_i)$ denotes the location of $H_i$ and $hc_i$ denotes the number of commuters who have the home location here but with the workplace located outside of this service scope.

A set of workplace location of cross-region commuters $W = \{wloc_j(x_j, y_j)\}$, in which $wloc_j(x_j, y_j)$ denotes the location of $W_j$ and $wc_j$ denotes the number of commuters who have the workplace location here but with the home located outside of this service scope.

To consider the home and workplace at the same time, a set of weighted locations representing the demand can be denoted as $P = H \cup W = \{P_k(loc_k, c_k)\}$ and the weight of each location is $c_k = hc_k + wc_k$. The objective is to select the center of the service scope according to the spatial distribution of demand $p$.

2. Methodology

2.1. Framework

The framework of the methodology is shown in Figure 3. In this paper, mobile phone data is used to observe the jobs-housing relationship and then help segment CB service in the urban agglomeration into hierarchical CB service scopes with CB stations. In brief, the methodology is as follows: First, using mobile phone data, the home and workplace of residents in the urban agglomeration are identified; the detailed data preprocessing of mobile signaling data and the location identification of home and workplace are shown in Supplementary Materials A. The jobs-housing relationship is then described by a jobs-housing flow network, counted and aggregated by 500 m × 500 m mesh grids. Second, an algorithm combining the method of community detection and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method is applied here to segment the CB service scope. The constructed jobs-housing flow network is iteratively segmented into hierarchical subcommunities with spatial clusters. The resulting hierarchical network structure represents the CB service scope segmentation in the urban agglomeration (see Algorithm 1). Then, the location selection method for CB station is applied to determine the location of CB station in each service scope at different levels. Finally, the hierarchical CB system is generated.

![Figure 3. The framework of the methodology.](image)

The iteration algorithm of generating a hierarchical CB system is the central part of the methodology, which is shown in Algorithm 1. In the beginning, the jobs-housing network in the whole study area is regarded as a single community. In each iteration, the algorithm includes the following three steps: (1) Network clusters segmentation: applying community detection to segment the community into densely connected subcommunities; (2) spatial cluster identification: applying DBSCAN algorithm to identify spatial clusters in subcommunities and remove the outliers without any spatial clusters (the remaining subcommunities are the CB service scopes); (3) location selection for CB station: based on the service scopes, to determine the location of CB station based on the cross-region commuting demand. The iteration will stop if there is less than or equal to one subservice scope in each service scope, implying that there is no possible segmentation on the existing hierarchical structure.

**Algorithm 1.** Hierarchical commuting oriented CB system service scope and station location selection algorithm.

Input: Jobs-housing relationship data
Output: A tree storing information of CB service scope and CB station location $T_s$

Algorithm:
1. Initialization: Set tree $T_s = \emptyset$ for storing communities on different level and append the whole study area as the only community of level 0 into $T_s$. Set $i = 0$ as the current level id.
2. While there is a community in the $i$th level of tree $T_s$
3. For community $c$ in the $i$th level of $T_s$
4. Construct the jobs-housing network $n_c$ of community $c$ base on the jobs-housing relationship data
5. Apply community detection method on $n_c$ and generate a subcommunity set $C$
6. For subcommunity $c_s$ in $C$:
7. Apply DBSCAN algorithm to identify spatial clusters in $c_s$
8. if $c_s$ do not contain any spatial cluster:
9. Remove $c_s$ from $C$
10. End if
11. End for
12. If there is more than one community in $C$:
13. For subcommunity $c_s$ in $C$:
14. Select CB station location for $c_s$
15. End for
16. Append the $i$th level subcommunity set $C$ into $T_s$ under the leaf node $c$
17. End if
18. $i = i + 1$
19. End for
End while

2.2. Constructing the Jobs-Housing Network

In this step, the methodology aims to extract the jobs-housing flow from the dataset and construct a network to represent the commuting demand. From the home and workplace identified from the dataset, the home places and workplaces are regarded as the origins (O) and the destinations (D). By aggregating the jobs-housing OD into mesh grids, we can obtain a network whose nodes are the cells of mesh grids and whose edges are present by the jobs-housing OD flow. The detailed methodology is as follow:

Given the identified home and workplace of each mobile subscriber, a weighted directed network $G(V, E, W)$ is created to represent the jobs-housing relationships in the study area. $V = \{v_1, v_2, \cdots, v_n\}$ is the set of nodes, representing $n$ grid cells in the study area. $E = \{e_{ij}\} \subseteq V \times V (i, j = 1, \cdots, n)$ is the set of edges representing the jobs-housing connections between grid cells. $e_{ij} = (v_i, v_j) \in E$ represents the jobs-housing connections of living in the grid cell $v_i$ and working in the grid cell $v_j$. $W = \{w_{ij}\}$ is the edge-weights. The value of $w_{ij}$ is defined as Equation (1):

$$w_{ij} = \begin{cases} q_{ij} & i \neq j \\ 0 & i = j \end{cases}$$  \hspace{1cm} (1)

where $q_{ij}$ denotes the number of subscribers who live in the grid cell $v_i$ and work in the grid cell $v_j$.

2.3. Service Scope Segmentation

In the service scope segmentation, the method consists of two steps. First, community detection is applied to the jobs-housing network to detect the network clusters with stronger intracommunity commuting demand. DBSCAN method is then applied to the communities to judge if it is forming any spatial cluster.
2.3.1. Network Clusters Segmentation by Community Detection

In the network analysis, a community is a collection of highly interconnected nodes [38]. The nodes belonging to different communities are sparsely connected. Community detection is the method to divide the network into densely connected communities. Modularity is an indicator between −1 and 1 that measures the density of links inside communities as compared to links between communities, which is widely used as the objective function in many community detection solution algorithms. It is defined as Equation (2):

\[
M = \frac{1}{2r} \sum_{ij} \left[ A_{ij} - \frac{k_i k_j}{2r} \right] \delta(c_i, c_j)
\]  

(2)

where \(A_{ij}\) represents the weight of the edge between \(i\) and \(j\), \(k_i\) is the sum of the weights of the edges attached to vertex \(i\), \(c_i\) is the community to which vertex \(i\) is assigned, the \(\delta\)-function \(\delta(u,v)\) is 1 if \(u = v\) and 0 otherwise, and \(r = \frac{1}{2} \sum_{ij} A_{ij}\).

To solve the problem of community detection, there are many algorithms. Here, the fast unfolding algorithm is adopted to decompose the network [39]. This algorithm is based on modularity optimization.

This algorithm includes the following two steps which are repeated iteratively until no increase of modularity is possible:

1. Modularity optimization: optimized modularity by allowing only local changes of communities;
2. Community aggregation: the identified communities are aggregated in order to build a new network of communities.

The advantage of adopting this algorithm is that the order of the nodes considered has no significant influence on the modularity, which ensures a stable result. The high computing speed of this algorithm makes it suitable for massive network analysis in this study. The fast unfolding toolkit provided in Python-igraph package is adopted in this study.

To measure the density of jobs-housing connections in each community, the graph density of the community is introduced here. It is defined as Equation (3):

\[
G = \frac{m}{n(n - 1)}
\]  

(3)

where \(n\) is the number of nodes and \(m\) is the number of edges in the community. The higher value of \(G\) indicates that the jobs-housing flow in the CB service scope is densely connecting the whole area into an integration.

2.3.2. Spatial Cluster Identification by DBSCAN

Since the community detection method is conducted without adjacency constraints, the subcommunities divided by the community detection can be spatially discrete. To avoid that, the DBSCAN algorithm is applied to identify spatial clusters for the community and rule out the communities without spatial clusters. DBSCAN is a classic density-based clustering algorithm [40]. This method performs clustering based on the density of the dataset in the spatial distribution. It can automatically determine the number of clusters and effectively process clusters of any shape. Given a threshold for the number of neighbors, minPts, and the radius, \(\epsilon\), objects with more than minPts neighbors within radius \(\epsilon\) are considered to be a core point. All neighbors within the \(\epsilon\) radius of a core point are considered to be part of the same cluster as the core point (direct density reachable). If any of these neighbors are also a core point, their neighborhoods are transitively included (density reachable). Noncore points in this set are called border points, and all points within the same set are density connected. Points that are not density reachable from any core points are considered to be outliers.

The selection of parameters minPts and \(\epsilon\) will influence the clustering results. Here, the data is aggregated into 500 m × 500 m mesh grids as the nodes in the network community. In order to
perform DBSCAN on network communities, the center of each grid is regarded as the position of the grid. The minPts are set as 10 points and the radius $\epsilon$ is set as 1 km, meaning that a grid will be considered as a core point if there are over 10 neighbors within 1 km. Figure 4 shows an example of the identification of spatial clusters in four network communities. Under this set of parameters, DBSCAN can distinguish if there exist any spatial clusters in the community.

![Figure 4. Spatial cluster identified by Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm.](image)

### 2.4. Location Selection Method for CB Stations

The location of CB stations is determined based on the spatial distribution of the demand $P_P\left(loc_k, c_k\right)$. Three methods are proposed as follows to generate the center of the spatial distribution of the demand $p$.

- **Weighted mean center (WA).**

  The weighted mean center is the weighted arithmetic average of the coordinates, it is a commonly used measure of central tendency and useful for tracking changes in the distribution or for comparing the distributions of different types of features [41]. The WA is formulated as Equations (4) and (5).

  \[
  x_s = \frac{\sum_k c_k x_k}{\sum_k c_k} \quad (4)
  \]
  \[
  y_s = \frac{\sum_k c_k y_k}{\sum_k c_k} \quad (5)
  \]

  where $c_k$ is the weight at location $k$. However, WA is easy to be influenced by outliers.

- **Weighted median center (MC).**

  The weighted median center is a measure of central tendency that is robust to outliers. It identifies the location that minimizes the weighted sum of the travel distances from the points in $p$. The objective function is as Equations (6) and (7):

  \[
  \min D = \sum_k c_k D_{s,k} \quad (6)
  \]
  \[
  D_{s,k} = \sqrt{(x_s - x_k)^2 + (y_s - y_k)^2} \quad (7)
  \]

  Dozens of papers have been written on the problem of solving this optimization with iterative algorithms: to choose a starting point and iteratively move the point closer to the median center. Comparing to the WA, the weight median center is more computationally expensive. Here, the algorithm proposed by Vardi and Zhang is applied to calculate the location of the weighted median center [42].

- **Weighted central element (CE).**
Weighted central element is the element in the dataset with the minimum of the weighted sum of the travel distances to the other elements. The difference between the central element and the median center is that the central element has to be one of the locations in \( p \).

3. Results and Discussion

3.1. Case Study

The study area in this study is the Yangtze River Urban Agglomeration in China, one of the largest urban agglomerations in the world with a population of over 70 million (Figure 5). It is one of the regions with the best economic development and the strongest comprehensive competitiveness in China. The study area is in the jurisdiction of two province-level administrations (Jiangsu Province and Shanghai), subdivided into one municipality directly under the central government (Shanghai) and nine prefecture-level cities. Since conflicts of interest between different districts can barely be moderated or coordinated, it is impossible to acquire complete and consistent datasets from different districts.

![Study area and its administrative division](image)

Figure 5. Study area and its administrative division.

Mobile signaling data is one type of mobile phone data provided by mobile operators, which was originally collected for billing and operational purposes. Compared with traditional datasets, mobile signaling data has the advantages of low collection cost, broad spatiotemporal coverage, and continuous tracking at an anonymous individual level.

The base transceiver station (BTS) is a piece of equipment that facilitates wireless communication between the mobile and the whole backbone network. Mobility is a distinct feature of the mobile system. The mobile continually monitors the signal strengths of the BTSs it can hear. As the mobile subscriber moves between different cell site range, the ongoing call or data session will be handed over from the BTS of the first cell to that of the next with no discernible disruption to the active data session. This process is called handover. As soon as a handover is induced, a piece of data will be
stored in the mobile signaling dataset, which is the basic principle of individual trajectory modeling using mobile signaling data. When modeling the individual trajectory, the space-time path of individuals is represented by the geographic location of the connected BTS and the timestamp of the data session conducted.

Moreover, once a call is placed or received, a text message is sent or received, or the mobile phone is switched on or switched off, the wireless communication will also be induced to transmit information between the mobile and BTS. As a result, these data session will also produce data records in the mobile signaling data, which constitutes a positive type of data redundancy to increase the continuity and smoothness of reconstructed individual trajectory.

Mobile signaling data used in this study were collected throughout the study area in November 2018, from a sample size of over 27 million users and with a daily data volume of about 80 million records. In the study area, BTSs are commonly spaced 500 m apart in the downtown areas and no more than 2 km apart in the suburban areas. The fields of mobile signaling data include the encrypted user ID, service time, service type, location of BTS, etc. In the study area, the dataset covered 9.96 million residents, accounting for 13.47% of the total population in the study area (Figure 6).

Figure 6. Spatial distribution of residents identified from the dataset. Aggregated into 500 m × 500 m mesh grids.
3.2. Result of Customized Bus Service Scope

The location of home and workplaces are identified using the mobile phone dataset collected in the Yangtze River Urban Agglomeration in China. The number of home and workplaces are counted and spatially aggregated by 500 m × 500 m mesh grids. A jobs-housing flow network is constructed to represent the commuting demand, with 111,438 nodes (each node is a grid) and 2,221,660 edges.

Applying the iteration algorithm of generating service scope, the algorithm iterates four times until the service scope cannot be further divided into spatially compact subservice scopes. The result indicates a four-level hierarchical service scope (Figure 7). The first-level CB service consists of 19 service scopes. Considering the size of service scopes, the first-level CB service mainly provides the service for intercity commuting demand. The level 2 includes 86 service scopes and mainly provides the CB service for intracity commuting demand. The service scopes of CB service at level 1 and level 2 can mostly cover the whole study area. The third-level CB service includes 134 service scopes and mainly provides the CB service for intracommunity commuting demand. Level 4 is the lowest level of CB service including 60 service scopes. The small-scale service scopes mainly serve the demand of intrasubcommunity commuting. The CB service scopes at level 3 and level 4 only cover the core areas of the urban agglomeration. Most of the third-level service scopes coincide with the metropolitan areas of Shanghai, Nanjing, Suzhou, Wuxi, and Nantong. The fourth-level service scopes consisting of the smaller communities mainly distribute around the essential parts in Shanghai, Nanjing, and Suzhou.
Figure 7. Hierarchical CB service scopes. The different colors represent different service scopes, CB stations are selected by the CE method: (a) level 1, (b) level 2, (c) level 3, and (d) level 4.

Comparing the service scopes with the administrative units, the service scopes in the urban agglomeration show the tendency of regional cohesiveness. Although the community detection method is dividing the region only according to the jobs-housing relationship, the resulting spatial divisions are so intertwined with the administrative boundaries, especially in the lower two levels of hierarchy. Apparently, in the current stage of regional integration, administrative boundaries still significantly affect the patterns of jobs-housing relationship.

At the same time, the partitioning of urban agglomeration has appeared the embryo of transjurisdictional integration, especially in the south of the Yangtze River. For example, Zhangjiagang in Suzhou and Jiangyin in Wuxi are two county-level cities under the jurisdiction of two prefecture-level cities. They are assigned together as a single service scope in level 1 but separated into 4 subservice scopes in level 2. The transjurisdictional integration in the CB service scope indicates that the integration of capital, labor, and industrial chains in the urban agglomeration has come into being. The traditional way to plan a CB system based on the administrative division is no longer suitable in the urban agglomeration.

The graph density \( G \) and the modularity \( M \) of each level of service scopes are shown in Figure 8. As the algorithm iteratively divides the communities at a higher level into smaller subcommunities at lower level, the indicator \( G \) gradually increases, whereas the modularity \( M \) gradually decreases. Graph density \( G \) is the indicator measuring the density of the edges in each community. The larger is the graph density \( G \) is, the larger the commuting demand between the service scopes is. Modularity \( M \) is the indicator measuring the strength of the division of a network into communities. The larger the modularity \( M \) is, the clearer the multicomunity structure exists in the jobs-housing network. As is shown in Figure 8, with the algorithm iteratively dividing the communities into smaller subcommunities, the indicator \( G \) gradually increases but the modularity \( M \) gradually decreases. This result indicates that compared to the larger division of service scope, the smaller division has denser intracommunity in a single service scope and also denser interconnection between service scopes. Generally speaking, the CB service provided in the service scopes at level 3 and level 4 mainly serves the strong demand for commuting within the metropolitan areas in the urban agglomeration. By comparison, the CB service provided in the first-level and second-level service scopes mainly satisfies the transregional commuting with lower frequency and less demand. The commuting connection in small cities or suburban areas of megacities is not strong enough to form spatial clustering areas. It is unaffordable to deploy CB service in smaller service scopes. Therefore, CB service at different levels can choose different operation strategies, including bus line design, bus vehicle type selection, and service level provided.
3.3. Result of Customized Bus Station Location Selection

In the three methods of CB station location selection, if the spatial distribution of the demand $p$ is in specific shape or with multiple clusters, the WA and MC can appear in the location without jobs or housing demand or even in the location impossible to access (in the lake or over the sea for instance). The CE is the center that can avoid the situation above since it is choosing the location with demand. Thus, in this paper, the CE is chosen as the location of CB stations.

Using the CE method to select the locations for customized bus stations, the result is shown in Figure 9. To show the hierarchical structure of CB network, leaf nodes are connected to the parent nodes by spanning tree. The area of Yangtze River Urban Agglomeration is divided into hierarchical tree structures, extends outwards from Level 1 stations to Level 4 stations and gradually covers the whole area.
Figure 9. The hierarchical structure of CB network. Leaf nodes are connected to the parent nodes by spanning tree.

Figure 10a shows the average travel distance between the demand \( p \) and the station location determined by the three methods proposed. Figure 10b shows the comparison of the number of population and job opportunities within 500 m and 1 km buffer areas surrounding the three types of stations.
In all levels of service scope, the station location determined by MC has the shortest average travel distance to the location of demand, whereas the station location determined by WA has the longest distance. The locations of WA are easy to fall into meaningless places with low population and job opportunities in its surrounding area. Comparing CE and WA, more population and job opportunities distribute in the 500 m buffer area of CE. However, when it comes to 1 km, WA surpasses CE. WA and CE usually distribute within a very close distance. In real-world application, choosing a method depends on the features of the areas and the density of service scopes. In the city center with denser and smaller service scopes, CE is better, whereas in suburban areas with larger service scopes, WA is recommended.

Figure 11 shows an example of a commuting trip served by CB network, i.e., a commuter with the origin and destination in Nanhui New City (Southeast part of Shanghai) and New Jiangwan City (Northeast part of the central urban area in Shanghai), respectively. These two areas are covered by level 3 service scopes under different level 1 parent nodes. The commuting path passes through level 3, level 2, and level 1 nodes to another level 1 node and then to level 2 and level 3 leaf nodes within its scope. Under this hierarchical form of node connection, the whole area is effectively connected.
Figure 11. Example of a commuting trip served by the CB network.

4. Conclusions and Future Prospects

Under the global background of transportation integration in the urban agglomeration, the commuting-oriented CB system is widely considered as a solution to deal with the growing commuting demand. However, a very few studies aimed at designing a hierarchical CB system corresponding to the comprehensive and hierarchical demand. A well-designed and dynamic system layout that matches the actual demand can directly affect the performance and the public acceptance of the CB system. The problem of generating a CB system layout involves the hierarchical CB station location selection from massive data, which is a coupling problem of hierarchical demand clustering and continuous Multi-facility Weber Problem. Meanwhile, the constant change of the demand also requires the dynamic and high computing speed of the method.

Therefore, this paper introduces a CB system layout design to simplify the problem and propose a method for the division of hierarchical service scopes and the location selection of CB stations. First, a method for home and workplace identification using mobile phone data is proposed. Based on the constructed jobs-housing relationship network, an algorithm is proposed to iteratively generate spatially consecutive service scopes with dense jobs-housing relationship using community detection and DBSCAN. Then, three methods of CB station location selection are proposed. A case study is conducted using the mobile phone dataset in the Yangtze River Urban Agglomeration in China. A four-level hierarchical CB system layout is generated by the methodology proposed in this paper. Then, both the spatial properties and network properties of the multi-level service scopes are analyzed. In the CB station location selection process, the proposed three methods are compared. As a result, the weighted central element method is proved to best suit the station location selection for its relatively reasonable CB station sitting and the averagely short distance to home and workplaces.

The proposed method has several application potentials:
• So long as the accurate and detailed data is provided, by applying the iterative algorithm, the service scopes can unlimitedly divide into smaller service scopes by changing the criteria, which enables us to choose an on-demand hierarchy structure.

• The iterative algorithm proposed in this paper is a powerful location selection tool, which is not limited to the scenario of CB system design. The methodology can be suitable for any OD flow dataset with uneven density in spatial cluster and have multiple application potentials (e.g., logistics center and rescue center location problem [43]).

In future studies, further analysis will have several directions. For example, the method proposed only generates the service scopes and location for CB stations without generating detailed route planning for the CB lines. In future studies, an optimization model of designing CB lines in detail can be built to enable a more comprehensive conclusion. The quantitative studies are significant for offering governments and CB operations for building infrastructure, policymaking, and potential market measurements.

Supplementary Materials A: The following are available online at www.mdpi.com/2071-1050/12/15/6203/s1.

Supplementary materials A: Data preprocessing of mobile signaling data and home & work location identification.

Author Contributions: Supervision, D.Y.; writing—original draft, Q.Y.; and writing—review and editing, W.L. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 71734004).

Conflicts of Interest: The authors declare no conflicts of interest.

References


