sustainability

MDPI

Article

An Analytical Model for the Many-to-One Demand Responsive Transit Systems

Di Huang, Weiping Tong *, Lumeng Wang and Xun Yang

Jiangsu Key Laboratory of Urban ITS, Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, School of Transportation, Southeast University, Nanjing 211189, China; dhuang2@seu.edu.cn (D.H.); lumeng.wang99@gmail.com (L.W.); yx_97@seu.edu.cn (X.Y.)

* Correspondence: wptong@seu.edu.cn

Received: 2 December 2019; Accepted: 29 December 2019; Published: 30 December 2019

Abstract: The demand-responsive transit (DRT) service is an emerging and flexible transit mode to enhance the mobility of the urban transit system by providing personalized services. Passengers can make advanced appointments through smartphone applications. In this paper, an analytical model is proposed for the many-to-one DRT system. The agency and user costs are approximated by closed-form expressions. The agency cost, which is also the operation cost, is approximated by the continuum approximation technique. A nearest-neighbor routing strategy is applied, whereby the vehicle always collects the nearest passenger waiting in the system. The Vickrey queueing theory is adopted as the basis for approximating each component of the user cost, which is composed of the out-of-vehicle and in-vehicle waiting times and schedule deviations, which also depend on the service quality of the DRT system. The results of the numerical experiment show that (1) the agency and user costs are influenced significantly by the demand density, and (2) the DRT operator cannot further decrease the operating cost by solely deploying larger vehicles.

Keywords: demand-responsive transit; analytical model; Vickrey equilibrium; optimal vehicle occupancy; many-to-one

1. Introduction

The construction of the on-demand transit system was widely recognized in the last decade as an effective policy that can solve the first/last mile connectivity between residential areas and public transit stations. Compared with traditional fixed-route transit, demand-responsive transit (DRT) provides flexible services owing to the on-demand schedules and dynamic deviations [1–4]. As an essential component of the multimodal transit system, the DRT usually acts as a conjunction by collecting passengers and carrying them to other transit stations (e.g., railway or fixed-route bus stations), which is an alternative travel option that is competitive with private automobiles.

One of the salient features of DRT is the elimination of access distance for transit users by providing door-to-door services where users only have to make travel requests and wait to be picked up. In general, two types of DRT system are widely applied in practice: fully flexible services and semi-flexible services. The fully flexible transit refers to the service with fully flexible routes and schedules, which are dependent on ad hoc demands. On the other hand, the semi-flexible service can be seen as the regular fixed-route transit with on-demand adjustments, such as route deviation [5]. This paper focuses on the fully flexible transit service.

Thanks to the advanced information and telecommunication technologies, travelers are allowed to take on-demand services or subscription bus services, especially in remote areas during peak hours [4,6]. An apparent trade-off is that transit users have to wait for the vehicle to pick them up and then stay in the vehicle, which may detour to pick up and drop off other users. Hence, the quality of

service could be evaluated by the waiting and in-service times, and the earliness/lateness with respect to their wished arrival time at their destinations [7–9].

The quality of the DRT service strictly depends on the operating cost of the agency. An apparent trade-off is that a large investment into transit operating resources (e.g., fleet size, crew, vehicle capacity) would increase the system service quality; conversely, the passengers may suffer from low service quality resulting from inadequate service capacity by reducing operating cost. For instance, passengers would tolerate higher delay in the system awaiting pickup services, longer in-vehicle travel time, and earliness/lateness resulting from the lack of punctuality if the passenger demand exceeds the existing operating capacity. In this regard, when transporting passengers, improving service quality must be balanced against minimizing operating costs by optimizing operation and management strategies in the DRT system, while the total costs of agency and passengers are minimized.

As a kind of flexible transit service, the routing and schedule of vehicles in the DRT system are responsive to the real-time passenger demand. One of the aims of DRT services is to improve the mobility in low-demand-density areas by providing door-to-door services. In this regard, the operating cost of DRT services becomes unavoidably expensive to maintain a considerable quality of service. Due to the limited operating resources, it is of great importance for the agency to use its resources as efficiently as possible and obtain an optimal balance between the expense of resources and quality of the DRT system. The major motivation of this study is to investigate how these critical parameters affect the system cost and the passenger's travel behavior in such a transit system.

To facilitate such a design, an analytical model is formulated to approximate the operating cost of the collection phase. This approach provides analytical tools for computing the travel distance through a set of demand points, which are assumed to be continuously distributed over a geographic region [8]. Compared with detailed techniques such as discrete mathematical programming methods, the analytical techniques formulate closed-form expressions based on the spatial density of demand rather than discrete points, which helps to reveal meaningful insights between essential components and usually reduces the computational burden [10].

From the perspective of transit users, they have to decide what time to make requests since they then wait at stops to be picked up. As aforementioned, an insufficient capacity of the DRT system would formulate a queue of service and lead to delays, as well as schedule deviations from users' wishes. The Vickrey [11] congestion theory is adapted to approximate the user cost incurred when the operating capacity is deficient to satisfy the demand. Transit users are assumed to be rational, who tend to minimize their own travel costs in their own trips by adjusting their request time to balance the trade-off between delay and schedule deviation, which leads to a user equilibrium (UE) condition, under which no one has an incentive to change their departure time to reduce their own travel cost. However, it is still possible to reduce the total system cost by optimizing the decisions variables (e.g., vehicle capacity, occupancy, feet size) during the DRT operation.

1.1. Literature Review

The DRT service was initially provided to extend the coverage of the traditional fixed-route network by providing door-to-door services and enhancing the mobility for certain groups of customers (e.g., the elderly and people with special needs), especially in remote areas with low demand densities. The practice the on-demand transit services emerged in the 1970s, i.e., the Dial-A-Ride service offered in the United States (US) [12]. With the help of interactive and intelligent information platforms (via smartphone applications or websites), various types of similar on-demand transit services gained increasing popularity worldwide, such as customized/subscription buses [4], dial-a-ride [13], personalized rapid transit [14], and shuttle buses [15]. The application of advanced planning and operational approaches, including intelligent dispatching and matching, flexible routing, and dynamic pricing, further makes the DRT an increasingly attractive alternative for personal vehicles and conventional transit systems [16,17]. Considering the continually increasing cost of on-demand

services, the modeling and estimation of the performance of DRT is a hot topic aimed at increasing the efficiency of the DRT system.

A great deal of research was conducted to investigate the factors influencing the performance of DRT systems from multiple aspects. Amirgholy and Gonzales [7] pointed out that the modeling approaches can be categorized in terms of the data availability, which are simulations [2,18–20] and mathematical methods [21,22]. The simulation method is applicable when it is difficult to develop closed-form expressions for the system's performance, which needs a large body of data including passenger demand and distribution (both spatially and temporally), fleet size, and vehicle trajectory. Dessouky et al. [20] conducted a study through simulation and investigated the level of service for specific strategies (e.g., zoning and time-window settings) conducted on the DRT system in Los Angeles, California. Chandra et al. [2] explored the optimal operating capacity, fleet size, and cycle length in El Cenizo, Texas, via a simulation study by using the data collected from a survey questionnaire.

A number of studies were also devoted to describing the DRT system analytically by formulating close-form mathematical expressions aimed at revealing the underlying relationships between essential variables. From the modeling point of view, the DRT system can be idealized as a many-to-many or many-to-one transportation network with respect to the demand characteristic and the loading/unloading strategy [22]. Daganzo [21] proposed an analytic model to measure the performance (e.g., average riding and waiting times) of the many-to-many DRT system for three dispatching strategies in terms of passenger demand and network attributes. It was found to fit the simulated result with a simpler and quicker calculation process. Fu [23] developed a new analytical model based on Daganzo's work [21] which could explain the complex relationship between system variables including fleet requirement, system capacity, and service quality. Rahimi et al. [17] proposed robust models to investigate the impacts of relevant exogenous variables on the agency's operating cost and validated the model with historical data.

The on-demand bus routing problem can be described as a traveler salesman problem (TSP) with pick-up and delivery, which was well studied by a large body of research aimed at improving the accuracy of the length of near-optimal tours. In practice, the passenger demand can be clustered both spatially and temporally. For instance, the destinations of passengers' morning commutes are usually concentrated at certain activity centers such as schools, hospitals, and transit hubs. Hence, the random many-to-many transportation network can be reduced to a many-to-few or many-to-one network via clustering and zoning techniques [23]. Daganzo [24] proposed a concise analytical expression to predict the travel distance of a fleet of vehicles by means of the "cluster first, route second" approach. The estimation model of the near-optimal length of the TSP tour in irregularly shaped areas was further developed in other studies [25,26]. Del Castillo [27] assumed that the set of demand points is distributed over a circular region. The TSP tour was constructed based on the optimal partition of the region, which was obtained by the continuous approximation of the demand points. Quadrifoglio et al. [15] introduced the rectilinear Hamiltonian path to the design of mobility allowance shuttle transit services. The relationship between the velocity along the primary direction and the demand was analyzed. Chandra and Quadrifoglio [2] developed an analytical model for the approximation of the near-optimal terminal-to-terminal cycle length of the DRT service. The objective function of this model was to maximize the system's level of service with respect to the cycle length.

Compared with conventional transit systems, a user's trip choice behavior in the DRT system is more complex because users have fully flexibility to make choices on their trips, including when to make trip requests, the specific pickup/delivery times, and whether to accept offered trips [4]. In rush hours, because of limited available capacity, the analysis of the congestion effect is of great importance, which would influence the user's travel behavior. In the 1960s, Vickrey [11] formulated the first tractable model of the dynamics of morning commuting congestion. The schedule delay cost was introduced to reflect the result of the congestion effect. In line with Vickrey's bottle model [11], Tian et al. [28] analyzed the equilibrium state of the morning commuting pattern on a many-to-one transit system considering the in-vehicle congestion effect and early/late arrival penalty. Amirgholy

and Gonzales [7] first considered the DRT system as a bottleneck where the Vickrey's [11] congestion theory was applied to capture the equilibrium condition and approximate user's costs. Each DRT user intends to minimize their own travel cost, including the waiting time to be picked up, the in-vehicle traveling time, and the schedule delays.

To sum up, even though there is a large body of research on analytical models of DRT systems, studies on the synthetic analysis of the DRT system from both the agency's and the user's perspectives are scarce. Additionally, most previous studies did not provide a comprehensive analysis of the user cost. For instance, the user's waiting time does not only refer to the time of waiting to be picked up. Because of the limited service capacity, the additional waiting time that a user waits to be assigned to an available bus should be taken into account.

1.2. Objectives and Contributions

The objective of this paper is to present an analytical model to investigate the relationship between essential variables and the agency and user costs. The agency cost is approximated by using the continuum approximation analysis. The passenger's travel behavior is assumed to follow Vickrey's equilibrium, where passengers have to decide what time to make the trip requests to DRT operators and consider the trade-offs between different cost components.

The contribution of this study is threefold. Firstly, closed-form expressions for the agency and user costs of the DRT system are proposed. Secondly, the user cost is considered comprehensively and approximated by the Vickrey's equilibrium where the new equilibrium state is derived for the DRT system. Thirdly, the experiment results show that the demand density is more influential than the area of the service region, and solely an increase in vehicle capacity cannot provide an efficient DRT system.

The remainder of this paper is organized as follows: Section 2 gives the analytical model for the DRT system. In Section 3, a series of numerical experiments are conducted. Finally, we conclude the paper with some remarks and perspectives.

2. Methodology

2.1. General Description of the Analytical Model

Consider a generic service region containing a hieratical transit network composed of two transit modes: the flexible on-demand transport service and the urban railway transit. The DRT serves as the feeder that collects passengers based on real-time requests and delivers them to railway stations. This operation mode is similar to the operating strategy for "many-to-one" systems [8], in which a fleet of homogeneous vehicles is assumed to travel in cycles, collecting passengers from their origins to railway stations. With perfect information (including the schedule information of both DRT and railway services), which can be easily obtained from the transit information platform provided by the agency (e.g., websites, smartphone applications), passengers are assumed to be able to decide their pick-up times aimed at catching the rail transit on time.

The service region is assumed to be characterized by its area, the number of railway stations, the spatial density of passengers over the region, and the depot location. Suppose a service region $\mathbf{A}$ with area A. Let $R = \{1, 2, \ldots, |R|\}$ denote the set of indices for railway stations in the service region and $j \in R$ be a particular station. We assume that each transit station serves a subregion of $\mathbf{A}$, which is known as the "catchment area" [29] that encompasses an area of potential passengers that would be willing to access this station. Each subregion is approximated by a rectangle with length and width, served by a fleet of homogeneous vehicles. Let $\mathbf{A}_j$ denote the subregion served by station j with area A_j. For simplicity, the service region is defined as fully covered by the subregion, that is, $\sum_{j \in R} A_j = A$. The user demand within the area is assumed to be spatially uniformly distributed within a given time interval of the day with a spatial density λ_j (number of users per unit area). Considering the full flexibility of the DRT service, stops of the DRT vehicle are according to the real-time passenger requests. The total number of stops, N_j, in area $\mathbf{A}_j$ is equal to the number of passengers, that is, $N_j = \lambda_j A_j$.

A fleet of identical vehicles with capacity Q (pax) is assumed to be allocated by the agency in the service area. Passengers make trip requests through the information platform to the agency to specify their desired pickup location and time. A nearest-neighbor routing strategy is adopted where each bus always picks up the nearest passenger who is waiting in the system. In the remainder of this section, a continuum approximation is firstly adopted to analyze the agency cost. Then, considering the queueing phenomenon, the Vickrey [11] queueing theory is adopted as the basis for approximating each component of the user cost.

2.2. Agency Cost

The agency costs in the DRT system include the expected total distance that vehicles travel in the network and the expected total fleet size needed in operation. The DRT network is constructed by a number of bus trips, each of which serves a railway station. More specifically, a bus trip begins at the depot and then travels to a subregion of size A_j to visit passengers in this area via a Hamiltonian path [30], while the number of passengers is obtained by $\lambda_j A_j$. In this study, it is assumed that the depot is located outside the service area **A**. Hence, it is common to decompose the expected distance of a DRT trip for a certain station, denoted as D, into three components: (i) the expected distance from the origin of the trip (e.g., the depot) to the first user in the subregion A_j, D_j^1; (ii) the expected distance from the first user to the last user within A_j, D_j^2; and (iii) the expected distance form the last user on the trip to the rail station, D_j^3. Hence,

$$D = D_j^1 + D_j^2 + D_j^3. \tag{1}$$

In continuum approximation analysis, the routing details within the vehicle service area are usually modeled approximately in terms of continuous variables, i.e., the spatial density [8]. Analytical models were developed in previous studies [26,31], which aimed at investigating the relationship between the tour length and the average number of passengers in service areas of different shapes based on the formulation in References [24,25]. For a spatially uniformly distribution of passenger demand, D_j^1 can be estimated by the expected distance from the depot to a passenger in $\mathbf{A}_j$ which is randomly located. Suppose that the distance from the origin of the trip (e.g., the depot) to the centroid of $\mathbf{A}_j$ is φ_{1j}. As developed by Campbell [31], D_j^1 can be calculated by the product of a factor, K_1, which is related to φ_{1j}, and the square root of A_j, that is, $D_j^1 = K_1(\varphi_{1j})\sqrt{A_j}$. D_j^2 indicates the sum of vehicle "peddling" distance with respect to the number of passengers in $\mathbf{A}_j$, which highly depends on the routing strategy the agency applied based on the specific circumstances of the serving area, such as demand density. Let φ_{2j} denote the distance from the centroid of $\mathbf{A}_j$ to station j. D_j^3 can then be obtained by $D_j^3 = K_1(\varphi_{2j})\sqrt{A_j}$.

In previous studies, the nearest-neighbor approach was widely adopted to approximate the expected passenger-to-passenger distance, in the analytical model of many-to-one or many-to-many dial-a-bus systems with respect to the passenger demand density [2,7,17,21]. It was proven to be a reasonable approximation and an efficient routing strategy. More specifically, after each pick-up service, the demand-responsive vehicle is routed to the nearest passenger who made the request. The approximation of the average distance to the nearest N_j uniformly distributed passengers over $\mathbf{A}_j$ can be written as

$$d_n(N_j) \approx K_2(N_j)\sqrt{\frac{A_j}{N_j}} = K_2(N_j)/\sqrt{\lambda_j}, \tag{2}$$

where $K_2(N_j)$ is the "peddling factor" defined by Daganzo [25] and Campbell [31], which is given in Table 1.

In sum, the total expected distance of a DRT trip for station j can be written as

$$D_j = K_1(\varphi_{1j})\sqrt{A_j} + N_j K_2(N_j)/\sqrt{\lambda_j} + K_1(\varphi_{2j})\sqrt{A_j}. \tag{3}$$

Table 1. Values of the peddling factor [31].

No. of Stops in $\mathbf{A}_j$, N_j	$K_2(N_j)$
1	0
2	0.73
3	0.68
4	0.63
5	0.60
≥6	0.57

Note that Daganzo [24] and Campbell [31] illustrated that the value of $K_1(\varphi_j)$ is related to three factors: (i) the distance metric (Euclidean or Manhattan), (ii) the distance between the origin/destination and the service area $\mathbf{A}_j$, and (iii) the shape of $\mathbf{A}_j$. In this paper, it is assumed that all distances are given by the Euclidean metric. $K_1(\varphi_j)$ can be determined from the equation developed by Vaughan [32].

$$K_1(\varphi_j) = \varphi_j \left[\frac{1}{\sqrt{A_j}} + \frac{\sqrt{A_j}}{8\pi\varphi_j^2} \right]. \tag{4}$$

Additionally, the agency should hold a sufficient number of vehicles to serve all passengers. Daganzo [21] illustrated that the minimum fleet size needed to satisfy the demand in an area depends on the average time that each passenger spends in the vehicle, which is composed of the stop time (e.g., pick-ups and drop-offs) and in-vehicle riding time. In the collection phase, the extra time required to serve an additional passenger includes the pick-up and drop-off times, as well as the vehicle riding time for the pick-up of the passenger. Hence, for a certain vehicle i, the length of the collection phase can be expressed as

$$C_i = n_{v,i}\left(b + \frac{d_{N,i}}{v}\right), \tag{5}$$

where b is the sum of pick-up and drop-off times, v is the average vehicle speed, and $d_{N,i}$ is the average distance traveled by the vehicle from a passenger to the next one when there are N passengers in the service area. $n_{v,i}$ is the number of passengers collected by vehicle i. The total service rate can be obtained by the sum of individual vehicles' service rates deployed in an area.

$$\mu = \sum_{i=1}^{M} \frac{n_{v,i}}{C_i}, \tag{6}$$

where M is the fleet size that the agency can use to adequately serve the demand by providing the total system service rate μ in this area. It is obvious that the vehicle collection time is an increasing function of the number of passengers waiting to be picked up. The attempt of collecting more passengers would inevitably result in a longer collection time, as well as an elongated average in-vehicle riding time for passengers on board, which has a negative effect on the system's level of service. Hence, one important design trade-off is between the level of service and the number of passengers each vehicle collects. Several studies illustrated that the best level of service is achieved when the number of passengers in each vehicle is equal [17,21]. Hence, Equation (6) can be rewritten as

$$\mu = \frac{M}{b + d_{N,i}/v}. \tag{7}$$

In the meantime, the optimal fleet size required in a certain area (say subregion j) can be obtained by $\left\lceil N_j/\widetilde{n}_{v,j} \right\rceil$, where $\widetilde{n}_{v,j}$ is the number of passengers for each vehicle collected in subregion j.

To sum up, the total operating distance of the agent can be represented as

$$AC\left(\left\{\widetilde{n}_{v,j}\right\}\right) = \sum_{j=1}^{R} M_j(\widetilde{n}_{v,j}) \cdot D_j(\widetilde{n}_{v,j}), \tag{8}$$

where M_j is the required fleet size in subregion j.

2.3. User Cost

In the DRT system, the agency provides a flexible door-to-door transportation service based on real-time users' requests. Experienced users who are assumed to have the full information of the system may evaluate their general costs including queueing delay and schedule deviations relative to their desired pick-up times, and they may maintain their trips by choosing their own requested pick-up times. Additionally, aimed at improving the operation efficiency, the agency prefers to elongate the route and deviate to collect more passengers. An unavoidable inconvenience would, however, be incurred for the passenger through the increased in-vehicle riding time [33]. Except for the queueing delay and schedule deviations, it is reasonable to consider the in-vehicle riding time as an irreducible component of the user cost.

Consider that, in a generic subregion j, the passengers are scattered throughout and wish to be picked up at the moment of their requests. Once the service rate is determined by the agency, it is broadcasted to all passengers to determine when they would make a request. Although making a request as early as possible might be a rational decision, the early boarding may cause other costs, e.g., the in-vehicle waiting time for the last passenger to board the vehicle. Each individual would wisely determine the request time to balance each component of their travel cost, including the waiting time in the queue out of the vehicle, the earliness penalty, the lateness penalty, and the waiting time for the departure of the vehicle. Thus, the total travel cost for a passenger arriving at time t is

$$\begin{aligned} C(t) = \; & \beta(\text{out-of-vehicle waiting time}) + e\beta(\text{earliness}) + l\beta(\text{lateness}) \\ & + w\beta(\text{in-vehicle waiting time}), \end{aligned} \tag{9}$$

where β (\$/time) is the cost rate that each passenger values in terms of the time in queues occurred out of the vehicle and in the vehicle. The cost rates of earliness time and lateness time are $e\beta$ and $l\beta$, respectively, where scalars e and l satisfy $e \leq 1 \leq l$ [34,35]. The cost rate of in-vehicle waiting delay is represented by $w\beta$. Furthermore, all passengers are assumed to be well-informed and rational, all of whom have an identical perception of each component of their costs.

The DRT system is assumed to be operated on a first-in-first-out order basis. In other words, no one can be picked up earlier upon making a request later. In an oversaturated system, the passenger demand rate will exceed the service rate because of the adequate operating capacity, which makes it impossible for vehicles to serve all passengers at the time they make requests. Rational passengers could adapt their request times by putting forward or postponing their requests to minimize their own travel costs. The passenger who boards the vehicle earlier has to stay in the vehicle, while the vehicle detours to pick up other passengers who requested the pick-up service. The cumulative result of these individual decisions would also lead to an equilibrium condition where no one has an incentive to change their request times. Considering these similarities, the passengers' request time choice behavior can be described by the Vickrey's equilibrium model [7,11,35]. The trade-offs of the passenger's travel cost lie in the tolerance of increased delay in waiting for pickups or longer in-vehicle time, as well as schedule deviations, which also depend on the service quality of the DRT system. In practice, the service quality or the performance of the DRT system is directly or indirectly influenced by many factors, especially the routing strategy adopted by the operator [7].

Figure 1 illustrates the cumulative result of passengers using the DRT system in the equilibrium condition. Let t_a denote the first passenger's request time and t_b denote the last passenger's request

time, which also are the starting and ending times of the queue in the DRT system of subregion j. $W(t)$ depicts the cumulative distribution of passengers' wished pick-up time. In this study, the passengers are assumed to have a common boarding time t^*, which is also illustrated by line EG. $D(t)$ is the cumulative pick-up curve, which describes the times when the passengers are actually picked up by the vehicle. The performance or the quality of service of the queueing system can be measured by the slope of $D(t)$, which is defined as the service rate (or effective system capacity, which indicates the number of passengers serviced per unit time) in Section 2.2 (see Equation (7)).

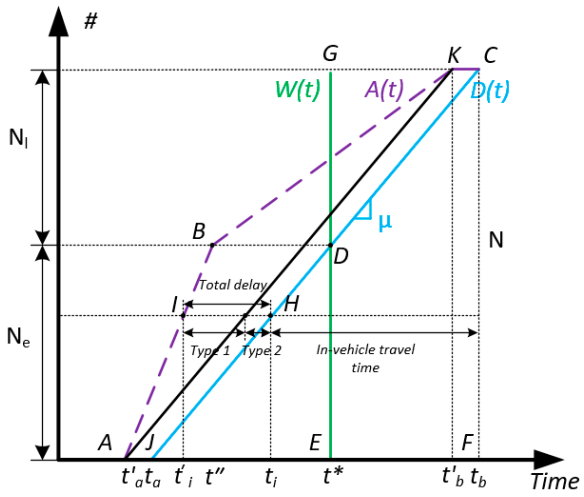


Figure 1. User equilibrium queueing diagram at a stop within a zone.

In the case where the vehicle detours over the region, the visiting sequence of passengers is scheduled as they enter the system, which can be seen as a queue that forms virtually in the DRT system. Suppose that a generic passenger is considered as the next one to be served, an extra waiting time is incurred because the vehicle has to travel from its current position to this passenger. In other words, two types of delay can be found: virtual queueing delay (termed as Type 1 delay) and actual vehicle traveling delay (termed as Type 2 delay). As illustrated in Figure 1, the first passenger who makes a request at t' would experience no virtual queueing delay but an unavoidable delay that indicates the minimum travel time required for the vehicle to receive the request and pick up the first passenger. Considering that the passengers are uniformly distributed within the subregion, this temporal gap, T_{gap}, can be approximated by the average passenger-to-passenger travel distance divided by the vehicle speed, that is, $T_{gap} = d_n/v$. Similarly, the last passenger has Type 2 delay only because they make the request at the time when there are no other passengers waiting in the system.

In the remainder of this section, we focus on the derivation of the UE passenger queueing pattern based on the bottleneck in an oversaturated DRT system in the subregion j. For notation simplicity, we consider the subregion j as a generic service region and omit the subscript j in all relevant notations.

Following Wardrop's first principle [36], it is assumed that, at equilibrium, passengers are subjected to the same total travel cost regardless of their request time for the service. Hence, the sum of schedule deviation and in-vehicle waiting costs for the first request is equal to the schedule delay cost of the last.

$$e\beta(t^* - t_a) + w\beta(t_b - t_a) = l\beta(t_b - t^*), \tag{10}$$

where the term of $\beta \cdot T_{gap}$ is subtracted from both sides of the above equation. Since all passengers are picked up during the time period (t_a,t_b), we have the following equation:

$$\mu(t_b - t_a) = N, \tag{11}$$

where N is the total number of passengers in the service region. Simultaneously solving Equations (10) and (11), we can obtain the starting and ending times of the queue.

$$t_a = \frac{\mu(e-l)t^* + (w-l)N}{\mu(e+l)}. \tag{12}$$

$$t_b = \frac{\mu(e-l)t^* + (w+e)N}{\mu(e+l)}. \tag{13}$$

As shown in Figure 1, assume that the passenger who is picked up by the vehicle at time t_i (Point H), makes a request at time t'_i, which means that the waiting time in the queue is $t_i - t'_i$. The passenger who is picked up punctually at his wished pick-up time at time t^* bears the longest waiting time of $t^* - t''$. Recalling the UE principles that passengers have the same costs, this yields

$$\begin{array}{ll} \beta \cdot T_{gap} + e\beta(t^* - t_a) + w\beta(t_b - t_a) & = \beta(t_i - t'_i) + e\beta(t^* - t_i) + w\beta(t_b - t_i) \quad (t_a \le t_i < t^*) \\ & = \beta(t^* - t'') + w\beta(t_b - t^*) \quad (t_i = t^*) \\ & = \beta(t_i - t'_i) + l\beta(t_i - t^*) + w\beta(t_b - t_i) \quad (t^* < t_i \le t_b). \end{array} \tag{14}$$

The solution of Equation (14) is shown as follows:

$$t'_i = (1 - e - w)t_i + (e + w)t_a - T_{gap}, \tag{15}$$

$$t'' = (1 - e - w)t^* + (e + w)t_a - T_{gap}, \tag{16}$$

$$t'_i = (1 + l - w)t_i + (e + w)t_a - (e + l)t^* - T_{gap}. \tag{17}$$

It is obvious that the cumulative of the request curve to the pick-up service, $A(t)$, is a piecewise linear curve (cumulative curve from point A to C) (see Figure 1). The remainder of this section provides the proof of this proposition.

Consider that an early passenger who is the n-th passenger makes a request before t^*. The total travel cost of the n-th passenger is

$$C(n) = \alpha\left[D^{-1}(n) - A^{-1}(n)\right] + \beta\left[t^* - D^{-1}(n)\right] + \eta\left[t_b - D^{-1}(n)\right], \tag{18}$$

where $A^{-1}(n)$ and $D^{-1}(n)$ are the inverse functions of $A(t)$ and $D(t)$, which denote the starting and ending queuing times of the n-th passenger, respectively. The equilibrium condition is obtained when no one has the incentive to change their arrival time at the queue as follows [37–39]:

$$\frac{\mathrm{d}C(n)}{\mathrm{d}n} = 0. \tag{19}$$

According to the property of the inverse function,

$$\frac{\mathrm{d}D^{-1}(n)}{\mathrm{d}n} = \frac{1}{dD(t)/\mathrm{d}t}. \tag{20}$$

By integrating Equations (18)–(20), we obtain

$$\frac{\mathrm{d}A^{-1}(n)}{\mathrm{d}n} = (1 - e - w) \cdot \frac{1}{dD(t)/\mathrm{d}t}. \tag{21}$$

It is obvious that both $A(t)$ and $A^{-1}(n)$ are linear functions. The slop of $A(t)$ is

$$\frac{\mathrm{d}A(t)}{\mathrm{d}t} = \frac{1}{1 - e - w} \cdot \frac{dD(t)}{\mathrm{d}t} = \frac{\mu}{1 - e - w}, \; t_a \le t_i < t^*. \tag{22}$$

Correspondingly, the slop of $A(t)$ during (t^*, t_b) is given by

$$\frac{\mathrm{d}A(t)}{\mathrm{d}t} = \frac{1}{1+l-w} \cdot \frac{\mathrm{d}D(t)}{\mathrm{d}t} = \frac{\mu}{1+l-w}, \; t^* \le t_i < t_b. \tag{23}$$

In contrast to previous papers where the total travel cost in the Vickery's equilibrium model was composed of queueing delay and schedule deviations only [7,11,36,39,40], the additional consideration of the in-vehicle delay cost in this paper would also result in two new equilibrium conditions. Firstly, the proportion of the early passengers, N_e, to the late passengers, N_l, is equal to

$$\frac{N_e}{N_l} = \frac{l-w}{e+w}, \tag{24}$$

and $N_e + N_l = N$. Secondly, the request curve in equilibrium is piecewise linear and must satisfy

$$\frac{\mathrm{d}A(t)}{\mathrm{d}t} = \begin{cases} \dfrac{\mu}{1-e-w}, & \text{for customers who arrives early} \\ \dfrac{\mu}{1+l-w}, & \text{for customers who arrives late.} \end{cases} \tag{25}$$

In the equilibrium condition, no one can reduce their travel time by making the request earlier or later. As a result, the critical passenger who is picked up at the time they wish would suffer the maximum delay (making request at point B).

$$T_C(N,\mu) = |BD| = \frac{(e+w)N_e}{\mu} + T_{gap} = \frac{(e+w)(l-w)N}{\mu(l+e)} + T_{gap}. \tag{26}$$

On the basis of the passenger queueing model in the DRT system, the approximation of user cost is presented in the remainder of this section. The passenger's total queueing delay (including both types of delay) is defined as the time difference between the time that the passenger makes a request and the time they are considered as the next one to be served, i.e., the horizontal distance between $A(t)$ and $D(t)$ in Figure 1. It is obvious that these two types of delay can be calculated separately. The Type 1 delay can be approximated by the area between $A(t)$ and line AK, which depicts the cumulative distribution of times that passengers are taken as the next ones to pick up. Additionally, passengers are assumed to experience the same Type 2 delay in that passengers are uniformly distributed in the service area. Accordingly, the total delay that all passengers experience waiting to be picked up can be approximated by the area between $A(t)$ and $D(t)$ in Figure 1.

$$ODC(N,\mu) = \beta\left(\frac{T_C}{2}N + \frac{T_{gap}}{2}N\right) = \frac{\beta N}{2}\left[T_C(N,\mu) + T_{gap}\right]. \tag{27}$$

Earliness and lateness are defined as the temporal difference between the actual and wished boarding times. Hence, as shown in Figure 2, the total earliness for all passengers can be approximated by the area between $D(t)$ and the wished curve $t = t^*$ from t_a to t^* (triangle DEJ), while the total lateness is approximated by the area between $D(t)$ and $t = t^*$ from t^* to t_b (triangle CDG). The approximations of total earliness and lateness are presented as follows:

$$\begin{aligned} TE(N,\mu) &= \frac{e\beta\mu}{2}(N_e)^2 = \frac{e\beta\mu}{2}\left(\frac{N(l-w)}{l+e}\right)^2 = \frac{e\beta\mu N^2(l-w)^2}{2(l+e)^2} \\ TL(N,\mu) &= \frac{l\beta\mu}{2}(N_l)^2 = \frac{l\beta\mu}{2}\left(\frac{N(e+w)}{l+e}\right)^2 = \frac{l\beta\mu N^2(e+w)^2}{2(l+e)^2}. \end{aligned} \tag{28}$$

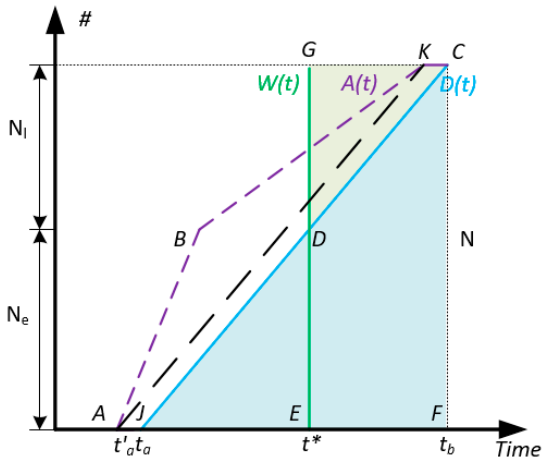


Figure 2. Illustration of different user costs.

By definition, the in-vehicle riding time for an individual passenger is measured by the difference between the time they are actually picked up and the ending time of the collection. As depicted in Figure 1, the total in-vehicle riding cost is proportional to the area of triangle *CFJ*, which can be approximated as follows:

$$IDC(N,\mu) = \frac{w}{2\mu}N^2. \tag{29}$$

In sum, the total travel cost of the passenger is confirmed to be the summation of the total waiting costs, including both out-of-vehicle and in-vehicle costs, and the total schedule deviations,

$$TTC(N,\mu) = ODC(N,\mu) + TE(N,\mu) + TL(N,\mu) + IDC(N,\mu) \tag{30}$$

2.4. Model Formulation

The identification of an efficient and attractive DRT system should satisfy a proper trade-off between the agency and user perspectives. As analyzed above, the goal of the agency is to determine the number of passengers, $\widetilde{n}_{v,j}$, collected by each bus in a specific subregion j with respect to regional characteristics. Once $\widetilde{n}_{v,j}$ is determined, a fleet of vehicles is deployed to the subregion and starts to collect passengers following a constant service rate μ_j, which is a monotonically nondecreasing function of $\widetilde{n}_{v,j}$ (see Equation (7)). Passengers who have perfect information regarding their travel cost with respect to μ_j will determine their request time to balance the trade-off between each component of their total travel cost. Specifically, the analytic model of the user cost can now be represented as a function of $\widetilde{n}_{v,j}$.

After all, the analytical model of the total system cost can be formulated as the weighted summation of the agency operation cost and the user cost with respect to $\widetilde{n}_{v,j}$, which considers the agency and the user points of view simultaneously; that is,

$$Z\left(\left\{\widetilde{n}_{v,j}\right\}\right) = \gamma_1 \cdot C \cdot AC\left(\widetilde{n}_{v,j}\right) + \gamma_2 \cdot TTC\left(\widetilde{n}_{v,j}\right), \tag{31}$$

where C is the vehicle operating cost (\$/veh h). γ_1 and γ_2 are weights that reflect the relative importance of the agency and user costs.

3. Numerical Experiments

In this section, the proposed analytical model is employed in a series of numerical experiments. Firstly, the model is employed on a hypothetical area composed of a set of railway stations and divided into subregions using the parameter values in Table 2. The impacts of changes in area or demand density (both singly and combinedly) on system costs are analyzed. The sensitivity analysis

is conducted to investigate how the key parameters affect the agency and user costs, as well as the passenger's travel choices.

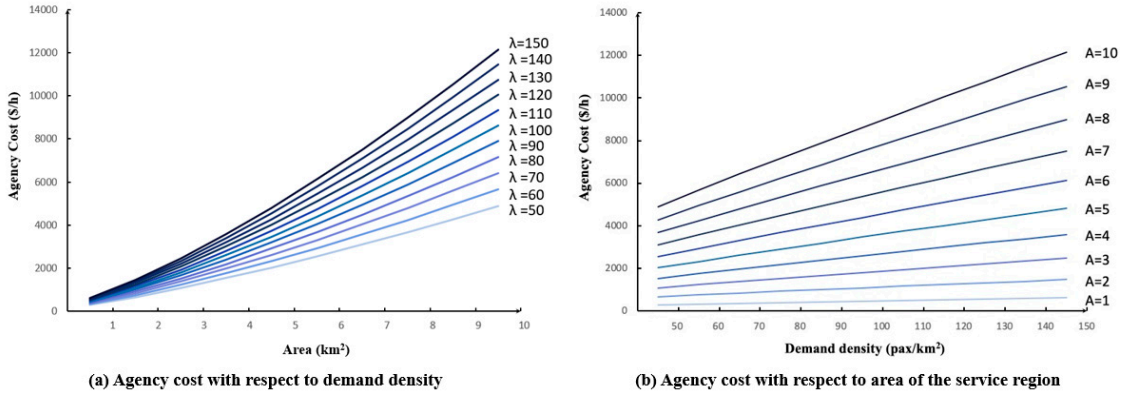
Table 2. Variable definitions.

Symbol	Definition	Baseline Value
A	Area of the service region (km^2)	-
b	Sum of pick-up and drop-off times (h)	0.08
C	Vehicle operating cost ($/veh h)	50
e	Scalar factor for earliness delay	0.5
l	Scalar factor for lateness delay	1.5
N	Number of passengers (pax)	-
N_e	Number of early passengers (pax)	-
N_l	Number of late passengers (pax)	-
Q	Vehicle capacity (pax)	30
v	Average vehicle speed (km/h)	30
w	Scalar factor for in-vehicle waiting delay	1.0
β	Cost rate of delay ($/h)	5
γ_1	The weight of agency cost	1
γ_2	The weight of user cost	1
λ	Demand density (pax/km^2)	-
μ	System service rate (pax/h)	-
φ_{1j}	Distance from the depot to the centroid of $\mathbf{A}_j$ (km)	-
φ_{2j}	Distance from the centroid of $\mathbf{A}_j$ to station j (km)	-

3.1. Effects of Area of the Service Region and Demand Density Changes

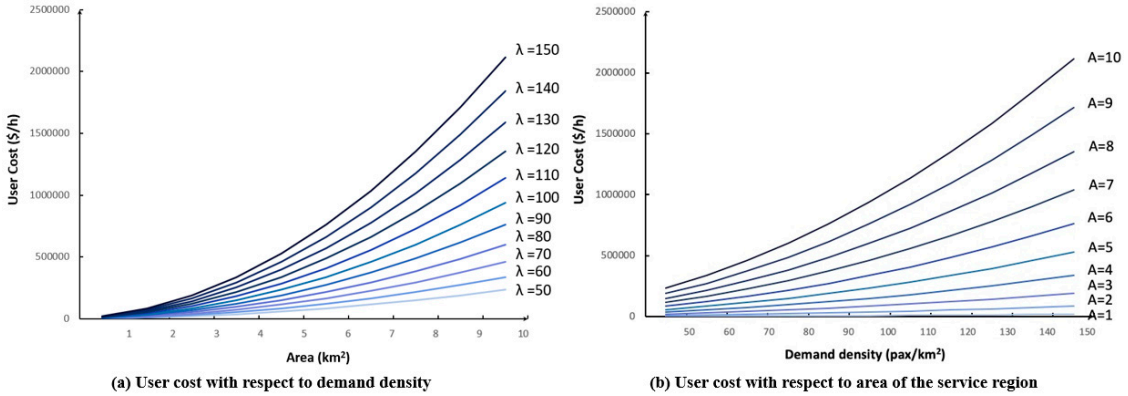
Figures 3 and 4 describe the effects of the regional area and the demand density on agency and user costs, respectively. Figure 3a shows that, if the DRT is operated in a small area, the influence of the demand density on agency cost is trivial. That is because, in such areas, the demand distribution is more concentrated, such as large residential communities, which results in the short transportation distance of DRTs. The influence of the demand density grows with the increase of the regional area, which is also illustrated in Figure 3b. In sum, this comparison shows that the demand density is more influential than the area of the service region. Given the system service level (i.e., service rate), the agency cost would increase inevitably with more passenger demand.

Figure 4 describes the influences of regional area and demand density on user cost, which shows that these two exogenous factors have greater impacts on the user cost than on the agency cost. Similar to Figure 3, the impact of demand density on user cost is not significant in a small service area. However, this impact increases rapidly with the growing of demand density. This increment mainly originates from two aspects: (1) the increase in out-of-vehicle waiting time (Type 1 delay) and (2) in-vehicle waiting time (Type 2 delay). In the next subsection, the changes among different components of user cost are discussed extensively.



(a) Agency cost with respect to demand density **(b) Agency cost with respect to area of the service region**

Figure 3. Effects of area of the service region and demand density on agency cost.



(a) User cost with respect to demand density **(b) User cost with respect to area of the service region**

Figure 4. Effects of area of the service region and demand density on user cost.

3.2. Effects of Vehicle Occupancy/Fleet Size Changes

As aforementioned, the vehicle occupancy, in other words, the number of passengers collected by each bus, is one of the essential variables that affect both the agency and the user costs, as well as the entire system's level of service. Figures 5 and 6 depict how the different components of system costs vary when the regulated occupancy changes. Figure 6 shows that both the agency cost and the total user cost drop significantly when the vehicle occupancy is under 14 (pax/vehicle). It indicates that the DRT service, despite being a flexible and personalized transit mode, still has the characteristics of a public transit system, i.e., economies of scale. When the regulated occupancy is less than four, the DRT service can be considered as a typical ride-sharing or taxi service, which needs a higher operation cost than buses. The agency cost becomes steady when the regulated occupancy is larger than 20, which means that the DRT operator cannot further decrease their operating cost by deploying vehicles with larger capacity.

Concerning the user costs (Figure 6), only the lateness decreases significantly with the increase in vehicle occupancy, while the Type 2 delay (i.e., the in-vehicle time) increases. That is because the vehicle with a larger capacity would pick up more passengers in a trip. However, according to Equation (6), the increase in occupancy increases the system's average service rate, which decreases the unit time of serving a passenger. Hence, passengers could make trip requests close to their desired time to decrease their travel costs.

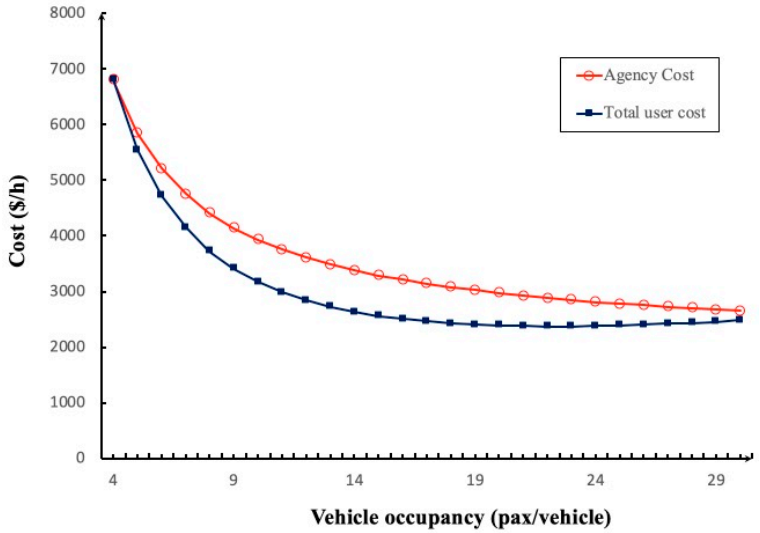


Figure 5. Agency and user costs versus the regulated vehicle occupancy.

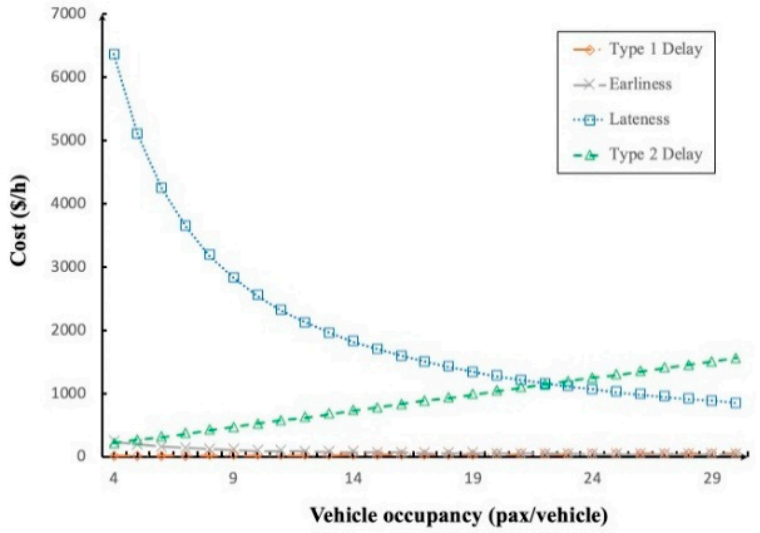
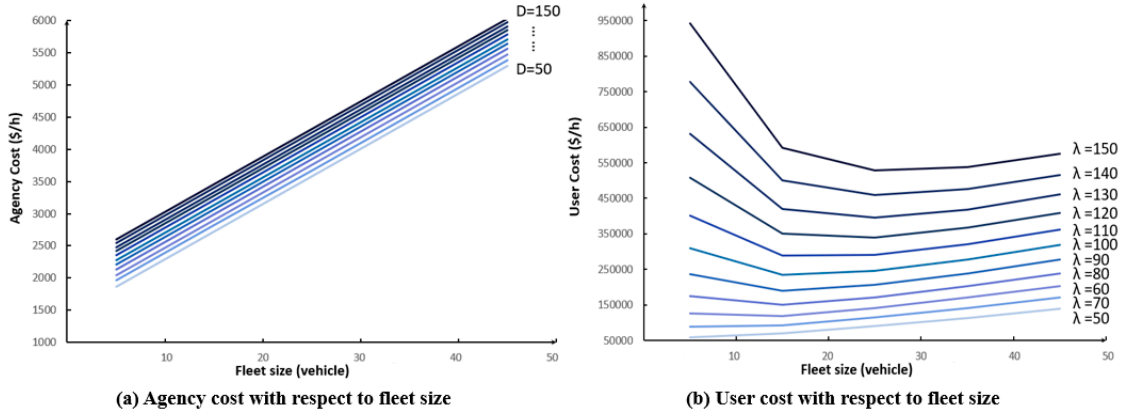


Figure 6. Different components versus the regulated vehicle occupancy.

As another essential attribute to the agency cost, the effect of required fleet size on the agency cost was further investigated. Figure 7a shows that the growth of the agency cost bears a good linear relationship with growth in fleet size for a certain demand density level. Regarding the user cost (see Figure 7b), the user cost drops sharply when the fleet size is lower than 20. When the fleet size is larger than 30, the decrease in user cost becomes slight. This indicates that dispatching more vehicle is only effective when the current fleet size is insufficient. Then, further increases in fleet size would not result in a significant decrease in user cost but an inevitable increase in agency cost.



(a) Agency cost with respect to fleet size **(b) User cost with respect to fleet size**

Figure 7. Effects of fleet size and demand density on user cost and agency cost.

4. Conclusions

The flexible and on-demand transit service is considered as an efficient way to address the first/last mile problem. In this paper, an analytical model for the DRT system was proposed and analyzed, and the closed-form expressions of the agency and user costs were presented. The agency cost, or the operation cost, is approximated by the continuum approximation technique, which is dependent on the average number of passengers allowed to be collected by each vehicle. The Vickrey queueing theory was adopted as the basis for approximating each component of the user cost. Two types of delay were considered in accordance with the characteristics of the DRT system, namely, out-of-vehicle and in-vehicle waiting times. The out-of-vehicle waiting time is of great importance in practice, and it measures the time period between the time a passenger makes the trip request and the time they are picked up. The results of the numerical experiment show that (1) the demand density is more influential than the area of the service region; (2) given a certain service area, the DRT operator cannot further decrease their operating cost by solely deploying larger vehicles; and (3) the increase in fleet size can efficiently reduce the user cost by decrease the waiting time when the number of dispatched vehicles is lower than 40. The further increase in fleet size was shown to be not cost-effective because the agency cost has a good linear relationship with the value of fleet size. The analysis results also provide several instructions for local DRT operators. The operation cost of the DRT system is highly sensitive to the service and demand level because it provides a personalized and high service level. Hence, a comprehensive analysis of both spatial and temporal distributions of user demand is necessary before providing DRT services. Considering the spatial and temporal heterogeneities of user demand, a zone-based operating strategy is preferred which can properly accommodate the demand distribution.

Several potential enhancements could be considered in future studies. Firstly, the proposed model can be extended to a many-to-many transit system by applying different routing strategies, such as alternating pick-ups and drop-offs or "collect first, deliver second". Secondly, the spatial and temporal stochasticity of passenger demand can be considered.

Author Contributions: Conceptualization, D.H.; methodology, D.H.; software, D.H.; validation, D.H., L.W., and X.Y.; formal analysis, D.H. and L.W.; writing—original draft preparation, D.H.; writing—review and editing, L.W. and X.Y.; visualization, D.H.; supervision, W.T.; project administration, W.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (No. 2018YFB1600900), the General Project (No. 71771050) and Key Project (No. 51638004) of the National Natural Science Foundation of China, and the Scientific Research Foundation of the Graduate School of Southeast University (No. YBPY1835).

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

References

1. Ciaffi, F.; Cipriani, E.; Petrelli, M. Feeder bus network design problem: A new metaheuristic procedure and real size applications. *Procedia Soc. Behav. Sci.* **2012**, *54*, 798–807. [CrossRef]
2. Chandra, S.; Quadrifoglio, L. A model for estimating the optimal cycle length of demand responsive feeder transit services. *Transp. Res. Part B* **2013**, *51*, 1–16. [CrossRef]
3. De Gruyter, C.; Currie, G.; Rose, G. Sustainability measures of urban public transport in cities: A world review and focus on the Asia/Middle East Region. *Sustainability* **2017**, *9*, 43. [CrossRef]
4. Huang, D.; Gu, Y.; Wang, S.; Liu, Z.; Zhang, W. A two-phase optimization model for the demand-responsive customized bus network design. *Transp. Res. Part C* **2020**, *111*, 1–21. [CrossRef]
5. Errico, F.; Crainic, T.G.; Malucelli, F.; Nonato, M. A survey on planning semi-flexible transit systems: Methodological issues and a unifying framework. *Transp. Res. Part C* **2013**, *36*, 324–338. [CrossRef]
6. Han, Z.; Chen, Y.; Li, H.; Zhang, K.; Sun, J. Customized bus network design based on individual reservation demands. *Sustainability* **2019**, *11*, 5535. [CrossRef]
7. Amirgholy, M.; Gonzales, E.J. Demand responsive transit systems with time-dependent demand: User equilibrium, system optimum, and management strategy. *Transp. Res. Part B* **2016**, *92*, 234–252. [CrossRef]
8. Daganzo, C.F. *Logistics Systems Analysis*; Springer: New York, NY, USA, 2005.
9. Bie, Y.; Xiong, X.; Yan, Y.; Qu, X. Dynamic headway control for high-frequency bus line based on speed guidance and intersection signal adjustment. *Comput. Aided Civ. Infrastruct. Eng.* **2020**, *34*, 4–25. [CrossRef]
10. Ansari, S.; Basdere, M.; Li, X.; Ouyang, Y. Advancements in continuous approximation models for logistics and transportation systems: 1996–2016. *Transp. Res. Part B* **2018**, *107*, 229–252. [CrossRef]
11. Vickrey, W.S. Congestion theory and transport investment. *Am. Econ. Rev.* **1969**, *59*, 251–260.
12. Ho, S.C.; Szeto, W.Y.; Kuo, Y.H.; Leung, J.M.; Petering, M.; Tou, T.W. A survey of dial-a-ride problems: Literature review and recent developments. *Transp. Res. Part B* **2018**, *111*, 395–421. [CrossRef]
13. Aldaihani, M.M.; Quadrifoglio, L.; Dessouky, M.M.; Hall, R. Network design for a grid hybrid transit service. *Transp. Res. Part A* **2004**, *38*, 511–530. [CrossRef]
14. Chebbi, O.; Chaouachi, J. Reducing the wasted transportation capacity of personal rapid transit systems: An integrated model and multi-objective optimization approach. *Transp. Res. Part E* **2016**, *89*, 236–258. [CrossRef]
15. Quadrifoglio, L.; Hall, R.W.; Dessouky, M.M. Performance and design of mobility allowance shuttle transit services: Bounds on the maximum longitudinal velocity. *Transp. Sci.* **2006**, *40*, 351–363. [CrossRef]
16. Tong, L.C.; Zhou, L.; Liu, J.; Zhou, X. Customized bus service design for jointly optimizing passenger-to-vehicle assignment and vehicle routing. *Transp. Res. Part E* **2017**, *85*, 451–475. [CrossRef]
17. Rahimi, M.; Amirgholy, M.; Gonzales, E.J. System modeling of demand responsive transportation services: Evaluating cost efficiency of service and coordinated taxi usage. *Transp. Res. Part E* **2018**, *112*, 66–83. [CrossRef]
18. Fu, L. Improving paratransit scheduling by accounting for dynamic and stochastic variations in travel time. *Transp. Res. Rec.* **1999**, *1666*, 74–81. [CrossRef]
19. Fu, L.; Teply, S. On-line and off-line routing and scheduling of dial-a-ride paratransit vehicles. *Comput. Aided Civ. Infrastruct. Eng.* **1999**, *14*, 309–319. [CrossRef]
20. Dessouky, M.; Ordóñez, F.; Quadrifoglio, F. *Productivity and Cost-Effectiveness of Demand Responsive Transit Systems*; California PATH Program, Institute of Transportation Studies, University of California at Berkeley: Berkeley, CA, USA, 2005.
21. Daganzo, C.F. An approximate analytic model of many-to-many demand responsive transportation systems. *Transp. Res.* **1978**, *12*, 325–333. [CrossRef]
22. Diana, M.; Dessouky, M.M.; Xia, N. A model for the fleet sizing of demand responsive transportation services with time windows. *Transp. Res. Part B* **2006**, *40*, 651–666. [CrossRef]
23. Fu, L. Analytical model for paratransit capacity and quality-of-service analysis. *Transp. Res. Rec.* **2003**, *1841*, 81–89. [CrossRef]
24. Daganzo, C.F. The distance traveled to visit N points with a maximum of C stops per vehicle: An analytic model and an application. *Transp. Sci.* **1984**, *18*, 331–350. [CrossRef]
25. Daganzo, C.F. The length of tours in zones of different shapes. *Transp. Res. Part B* **1984**, *18*, 135–145. [CrossRef]

26. Gaboune, B.; Laporte, G.; Soumis, F. Expected distances between two uniformly distributed random points in rectangles and rectangular parallelpipeds. *J. Oper. Res. Soc.* **1993**, *44*, 513–519. [CrossRef]
27. Del Castillo, J.M. A heuristic for the traveling salesman problem based on a continuous approximation. *Transp. Res. Part B* **1999**, *33*, 123–152. [CrossRef]
28. Tian, Q.; Huang, H.J.; Yang, H. Equilibrium properties of the morning peak-period commuting in a many-to-one mass transit system. *Transp. Res. Part B* **2007**, *41*, 616–631. [CrossRef]
29. Ceder, A. *Public Transit Planning and Operation: Theory. Modeling and Practice*; Elsevier: Oxford, UK, 2007.
30. Franceschetti, A.; Jabali, O.; Laporte, G. Continuous approximation models in freight distribution management. *Top* **2017**, *25*, 413–433. [CrossRef]
31. Campbell, J.F. One-to-many distribution with transshipments: An analytic model. *Transp. Sci.* **1993**, *27*, 330–340. [CrossRef]
32. Vaughan, R. Approximate formulas for average distances associated with zones. *Transp. Sci.* **1984**, *18*, 231–244. [CrossRef]
33. Fu, L. Planning and design of flex-route transit services. *Transp. Res. Rec.* **2002**, *1791*, 59–66. [CrossRef]
34. Small, K.A. The scheduling of consumer activities: Work trips. *Am. Econ. Rev.* **1982**, *72*, 467–479.
35. An, S.; Cui, N.; Li, X.; Ouyang, Y. Location planning for transit-based evacuation under the risk of service disruptions. *Transp. Res. Part B* **2013**, *54*, 1–16. [CrossRef]
36. Wardrop, J.G.; Whitehead, J.I. Some theoretical aspects of road traffic research. *Proc. Inst. Civ. Eng.* **1952**, *1*, 767–768. [CrossRef]
37. Arnott, R.; De Palma, A.; Lindsey, R. Economics of a bottleneck. *J. Urban Econ.* **1990**, *27*, 111–130. [CrossRef]
38. Lindsey, R. Existence, uniqueness, and trip cost function properties of user equilibrium in the bottleneck model with multiple user classes. *Transp. Sci.* **2004**, *38*, 293–314. [CrossRef]
39. Zhang, X.; Zhang, H.M.; Li, L. Analysis of user equilibrium traffic patterns on bottlenecks with time-varying capacities and their applications. *Int. J. Sustain. Transp.* **2010**, *4*, 56–74. [CrossRef]
40. Daganzo, C.F. The uniqueness of a time-dependent equilibrium distribution of arrivals at a single bottleneck. *Transp. Sci.* **1985**, *19*, 29–37. [CrossRef]