



Optimal design of sustainable transit systems in congested urban networks: A macroscopic approach



Mahyar Amirgholy^a, Mehrdad Shahabi^b, H. Oliver Gao^{a,*}

^a School of Civil and Environmental Engineering, Cornell University, Ithaca, NY 14853, United States

^b Department of Civil and Coastal Engineering, University of Florida, Gainesville, FL 32611, United States

ARTICLE INFO

Article history:

Received 27 November 2016

Accepted 16 March 2017

Available online 25 May 2017

Keywords:

Transportation system sustainability

Public transit systems

Macroscopic fundamental diagram

Continuum approximation

System optimization

ABSTRACT

Mass transit is a key component of a sustainable transportation system in urban networks. In this research, we propose a continuum approximation model to optimize the line spacing, stop spacing, headway, and fare of the transit system by minimizing a linear combination of (1) users generalized cost, (2) agency operating cost, and (3) external cost of the emission in the urban region. The design of the transit system can be optimized by minimizing the total cost of the transportation system in three different network allocation scenarios: (i) mixed network (Bus), (ii) dedicated lanes (BRT), and (iii) parallel networks (Metro).

© 2017 Published by Elsevier Ltd.

1. Introduction

Rapid growth in urban travel demand in large cities has heightened the need for alternative modes of transport that are socially beneficial, economically justifiable, and environmentally friendly. In populated urban regions, a transit system can play a prominent role in improving the sustainability of transportation by alleviating the unpleasant impacts of congestion across the network. Providing a quality of service competitive with the automobile mode can convince a considerable portion of users to leave their private cars at home and use the public transit system for their trips in the network. A competitive transit system can improve user mobility and reduce the emission of air pollutants in the network by decreasing the number of circulating vehicles and increasing the average speed of traffic flow. However, the share taken by a transit system from the network travel demand largely depends on the quality of service it offers to users. Meanwhile, improvement in the service quality of the transit system may entail a significant rise in the operating cost of a system. Therefore, it is of great importance to strike a balance between the operating cost and service quality of a transit system by minimizing the social, economic, and environmental costs of the system in the design of a sustainable transit system.

The network-wide transit design problem is introduced in Holroyd (1967) that proposes an analytical model for optimal headway and line spacing in a grid transit network by minimizing the summation of operating cost and user travel time. The classic analytical model of a transit system has also been extended for different types of network structures in the literature. Byrne (1975) designs a radial system with optimal line spacing and headway using a polar coordinate system. The radial design of the transit network has been further elaborated by adding rings to the radial system in Vaughan (1986). Byrne and Vuchic (1972) and Wirasinghe et al. (1977) also analytically model the operation of a system of corridors with feeder-bus service in grid and radial transit networks. The optimal design of a hub-and-spoke system with minimal operat-

* Corresponding author.

E-mail address: hg55@cornell.edu (H.O. Gao).

ing cost is also discussed in Newell (1979). The optimal network structure, operating characteristics, and fare pattern of transit systems have been further elaborated in recent studies (Yu et al., 2015; Bagherian et al., 2016; Chebbi and Chaouachi, 2016; Huang et al., 2016).

The stochasticity in the demand and supply sides of a transit system has been studied in the literature as well (Chien et al., 2001; Daganzo and Smilowitz, 2004). To optimize the infrastructure of a transit system with elastic demand, Daganzo (2012) maximizes the social welfare of the system with multiple classes of users. Promoting the efficiency of a transit system by improving the mobility of the feeder service and land-use pattern around the trunk sections has been studied in Sivakumaran et al. (2014). Amirgholy and Gonzales (2016) also optimizes the operation of demand responsive transit systems with time-dependent demand by minimizing the summation of the generalized cost of users and the operating cost of the service for agency costs.

The role of the transit system (as an alternative to automobile mode) in reducing congestion in urban regions has been widely investigated in the literature. Daganzo (2010a) proposes a design for a transit service with a competitive level of accessibility to the automobile at a reasonable cost by optimizing the headway and structure of a hybrid hub-and-spoke system with two levels of coverage in the central and peripheral districts of a grid network with uniform demand distribution. The concept of the hybrid system has been generalized in Estrada et al. (2011) to design a high-performance transit system for the city of Barcelona. Nourbakhsh and Ouyang (2012) also proposes an alternative flexible-route transit system with hybrid structure for low demand areas. The hybrid model has been recently extended in Badia et al. (2014) for the ring-and-radial network structure using a polar coordinate system. Chen et al. (2015a) also adopts the hybrid concept to develop two continuum approximation models for grid and ring-and-radial structures in the transit network that can be solved numerically for real size problems. The long-term effects of designing a ring-and-radial network structure for light rail transit in urban regions have been studied in Saidi et al. (2016).

Recent studies have taken steps towards accounting for the interaction between automobiles and the transit system in bimodal urban networks. Guler and Cassidy (2012) shows that the delay that users experience in arterial networks can be significantly reduced when the available capacity of the network is optimally shared between the automobile and transit modes. The strategy of dedicating separate lanes to the transit system to improve user mobility in the network is also investigated in Gonzales and Daganzo (2012) by extending Vickrey's (1969) congestion theory to model the user equilibrium condition between the automobile and transit modes in a bottleneck over the morning peak period. To optimize the bimodal system, a dynamic pricing strategy is proposed for both automobile and transit modes that makes it possible to eliminate the delay in the bottleneck. In another paper, these authors employ the bottleneck model to compare the equilibrium condition between the automobile and transit modes over the morning and evening peak periods (Gonzales and Daganzo, 2013). Building upon these papers, Daganzo (2013) proposes a day-long pricing strategy for both automobile and transit modes that can optimize the system by minimizing the generalized cost of commuters. Tian et al. (2013) and Xiao et al. (2015) also adopt time-dependent credit schemes for managing congestion in a bi-modal system with heterogeneous user preferences.

The idea of using the macroscopic fundamental diagram (MFD) for optimal allocation of network capacity to different modes of transport is shaped in Gonzales et al. (2010) and developed in Gonzales and Daganzo (2012). In large urban regions (neighborhoods), the macroscopic relationship between the aggregated traffic variables of the network, i.e. speed, flow, and density, can be represented by the MFD of the region. The MFD model is analytically developed in Daganzo (2007) and Daganzo and Geroliminis (2008), and empirically measured for the city of Yokohama in Geroliminis and Daganzo (2008). Properties of the MFD largely depend on characteristics of the network and the distribution of demand in the region. Recent research on the macroscopic relationship between the traffic variables of the network shows that the scatter and the shape of the MFD can be significantly affected by the spatial distribution of congestion in the network. In fact, heterogeneity in the spatial distribution of congestion can slow down traffic flow in the network, and such effects can even be magnified by a rise in the vehicle density of the system (Mazloumian et al., 2010; Geroliminis and Sun, 2011; Gayah and Daganzo, 2011; Mahmassani et al., 2013). In real size urban networks where the distribution of congestion is naturally not homogenous, the region can be partitioned into a reasonable number of homogenous clusters in order to derive well-defined (low-scattered) MFDs for the subregions of the network (Ji and Geroliminis, 2012). The effects of heterogeneity in the spatial distribution of the congestion in the network on the scatter, shape, and accuracy of the MFD have also been examined using region-based and subregion-based MFD models (Ramezani et al., 2015; Xie et al., 2016). The macroscopic relationship between the traffic variables in real urban networks can be approximated by estimating a limited number of observable parameters (Laval and Castrillón, 2015). Alternatively, the MFD of the region and the average vehicle density of the network can be estimated using low penetrated probe data from the network (Gayah and Dixit, 2013; Ji et al., 2014; Leclercq et al., 2014; Nagle and Gayah, 2014; Du et al., 2015a). This information can be crucial for designing real-time traffic control strategies in urban networks (Keyvan-Ekbatani et al., 2012; Haddad and Geroliminis, 2012; Geroliminis et al., 2013; Haddad et al., 2013; Gayah et al., 2014; Haddad and Shraiber, 2014; Du et al., 2015b; Ramezani et al., 2015). In addition to applications for the design of real-time control strategies, the MFD model can be applied in evaluating the performance of the transportation system for planning and design purposes.

In this paper, we make a use of the MFD model to capture the effect of operating a transit system on the average speed and outflow of the urban network. The objective of the research is to design a sustainable transit system with a quality of service competitive with the automobile by minimizing the social, economic, and environmental cost of the transportation system. To cope with the complexity of the problem, we put the focus of the paper on aggregating the behavior of users and the performance of the system by ignoring the fine details of the problem to develop a continuum approximation model that

can be used to answer big picture questions in large urban networks. In spite of such simplifications, the model remains accurate and robust as explained in [Daganzo \(2005\)](#). The continuum approximation model proposed in this paper simultaneously optimizes the network structure (stop spacing and line spacing) and the operating characteristics (headway and fare) of the transit system by minimizing a linear combination of the generalized cost that users experience in their trips, the operating cost of the transit system for the agency, and the external cost of emissions in a homogenous urban network.¹ The model is first developed for a grid network, and then extended for a ring-and-radial network structure by adapting the components of the costs using a polar coordinate system. In this model, the fixed travel demand of the network is assumed to have a general S-shaped cumulative distribution over the study period, which is uniformly distributed over the region. Identical users of the transportation network are assumed to have the flexibility to choose between automobile and transit modes for their trips in the network. Lines of the transit system lie along the arterial network and operate in both directions, while the characteristics of the transit system remain fixed all over the region over the peak. The average length of the trips in the network is assumed to be equal for both modes of transport, while transit users can transfer between lines to get to their destinations. To split the demand between the modes, we make use of a binomial logit model to determine the share of the automobile and the transit modes from total demand based on the average generalized cost of these modes for their users. The generalized cost that automobile users experience over the morning peak in the congested network is also approximated using a bottleneck model derived by combining [Vickrey's \(1969\)](#) congestion theory with the MFD. We also develop an analytical model for the transit system in which the effect of the congestion on the service quality and performance of the transit system is taken into account using the MFD model. This analytical model is then employed to approximate the generalized cost that transit users experience in their trips. The cost of constructing the required infrastructure and operating cost of the transit system for agency is also approximated as the summation of the costs associated with acquiring and maintaining the fleet, vehicles miles traveled (VMT), vehicle hours traveled (VHT), and infrastructure construction for the transit system minus the fare revenue. The external cost of air pollutants emitted by automobiles and transit vehicles is also calculated using a VMT-based fuel consumption model that is extended for the congested networks by approximating the average speed of the vehicles using the MFD model. To optimize the system, we consider three prevalent scenarios of network allocation to the transit system. In the first scenario, we consider the case where automobiles share the arterial network with transit vehicles, e.g. a conventional bus system. The second scenario considers the operation of the transit system with dedicated lanes in the arterial network, e.g. a Bus Rapid Transit (BRT) system. The third scenario also assumes that the transit system operates in a separate network parallel to the arterial network, e.g. a Metro system. The continuum approximation model proposed in this paper can be also extended for designing multimodal transit systems in large urban networks with multiclass users by employing a nested logit model for the modal split of the users. The optimal designs of the transit system in different network allocation scenarios are compared in a numerical example. The key contribution of this research is to develop a continuum approximation model that can be used to obtain the primary design guidelines of the sustainable transit systems in congested urban networks. The preliminary strategies and guidelines can be further elaborated to develop practicable plans in the fine-tuning step using detailed numerical techniques (see [Daganzo, 2010b; Daganzo et al., 2012](#)).

The rest of the paper is organized as follows: Section 2 reviews the MFD as a primary modeling tool used in this research. In Section 3, we analytically approximate the components of the total cost of the transportation system in a grid urban network. Section 4 presents the sustainable transit design problem in three different network allocation scenarios. In Section 5, we extended the model developed in Sections 3 and 4 for a ring-and-radial network structure. A numerical illustration of the proposed model is also provided in Section 6. Lastly, conclusions of the paper are summarized in Section 7.

2. Macroscopic fundamental diagram and network exit function

The macroscopic fundamental diagram (MFD) is introduced in recent studies to model the interrelationship between the speed, v (s/m), flow, q (veh/sec.lane), and density, k (veh/m.lane) of the network in large urban regions (neighborhoods). Research on observed traffic data from the city of Yokohama network ([Geroliminis and Daganzo, 2008](#)) and the results of microsimulations of the traffic flows in the San Francisco ([Geroliminis and Daganzo, 2007](#)) and Nairobi ([Gonzales et al., 2011](#)) networks indicate that the traffic flow in the network rises to its maximum feasible value (network capacity) as the density of the network increases in the uncongested state of the network (off the peak). Meanwhile, the average speed of the network remains close to the free flow speed of the network ([Geroliminis and Daganzo, 2008; Gonzales et al., 2010](#)), and can be assumed constant off the peak ([Gonzales and Daganzo, 2012; Haddad and Geroliminis, 2012](#)). However, in the hypercongested state condition of the network (over the peak), the network flow gradually decreases back to zero, with further rise in the density of the network as the system moves towards a complete gridlock, as plotted in [Fig. 1a](#). The relationship between speed, flow, and density in the network is also determined by the macroscopic flow equation, $q = kv$.

Having the macroscopic relationship between flow and density in a network, the relationship between outflow (number of vehicles that exit or park inside the region per unit of time) and vehicular accumulation in the region (number of vehicles circulating inside the region) can be represented by the network exit function (NEF) of the region. For this purpose, the MFD of the network, $Q(k)$, can be rescaled to approximate the network outflow, μ (veh/sec), as a function of number of the vehicles circulating inside the region, \bar{n} (veh) ([Daganzo, 2007; Gonzales and Daganzo, 2012](#)). In large urban regions, network out-

¹ Large heterogeneous networks can be partitioned into a number of homogenous sub-regions with well-defined MFDs ([Ji and Geroliminis, 2012](#)).

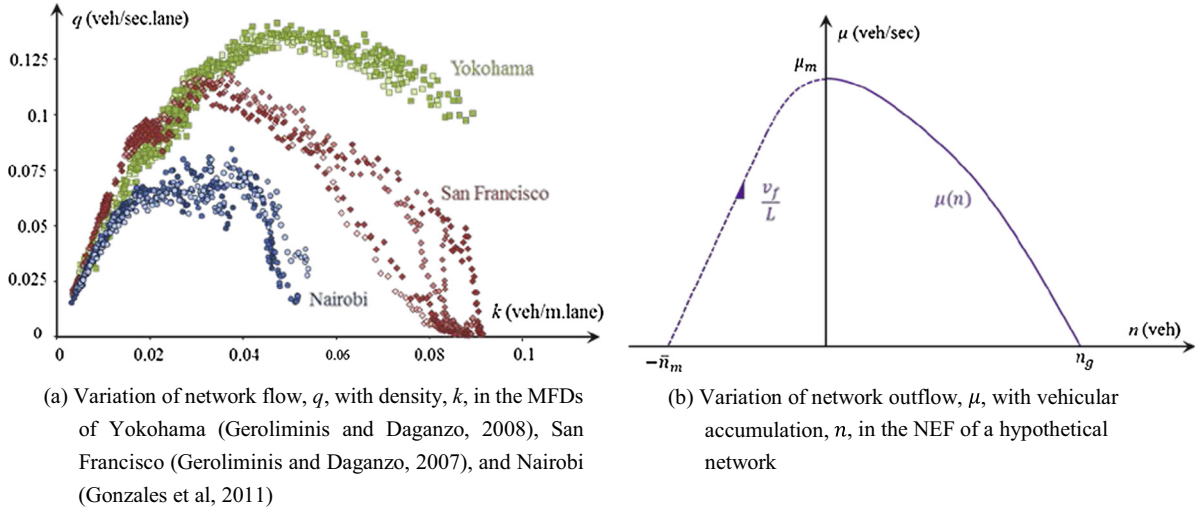


Fig. 1. Macroscopic fundamental diagram (MFD) and network exit function (NEF) of an urban network.

flow reaches its maximum feasible value when the number of vehicles circulating inside the network is at its optimal value, $\bar{n} = \bar{n}_m$. Here, we refer to the maximum feasible outflow of the network as the capacity of the network, μ_m . So, during the peak period where network capacity is insufficient to meet the demand, vehicular accumulation of the network exceeds its optimal value, $\bar{n} > \bar{n}_m$, and network outflow drops below capacity, $\mu < \mu_m$. Therefore, the outflow of the network over the peak period can be approximated just using the declining part of the NEF. Accordingly, we can just consider the declining part of the MFD model for approximating the outflow of the network over the peak period by laterally shifting the vertical axes (μ) to the turning point of the curve, $n = \bar{n} - \bar{n}_m$. As a result, the outflow of the network at each point in time can be approximated as a function of the normalized accumulation of the system by rescaling the declining part of the MFD as below:

$$\mu(n) = \frac{A_N}{L} Q\left(\frac{n + \bar{n}_m}{A_N}\right) \quad (1)$$

where A_N (m.lane) and L (m) denote the total area of the network and the average length of the trips in the network, respectively. Given the MFD of the network, the average speed of the automobiles in the network, v , can be also derived using the macroscopic flow equation as follows:

$$v(n) = \frac{\mu(n)}{n + \bar{n}_m} L = \frac{A_N}{n + \bar{n}_m} Q\left(\frac{n + \bar{n}_m}{A_N}\right) \quad (2)$$

In the remainder of this paper, we refer to n as the vehicular accumulation of the system. Fig. 1b depicts the declining part of the NEF for a hypothetical network.

3. Generalized cost of the transportation system

In designing a sustainable transit system, it is critically important to strike a balance between the social, economic, and environmental costs of the transportation system by minimizing the total generalized cost of the system in designing a sustainable transit service. To this end, the total cost of the transportation system (C), can be approximated as a linear combination of the generalized cost that users experience in the network (T_U), the construction and operating cost of the transit system for the agency (C_T), and the external cost of emissions (E_N) as below:

$$C = \beta_T T_U + \beta_C C_T + \beta_E E_N \quad (3)$$

where β_T , β_C , and β_E represent the relative importance factors of user generalized cost, the agency operating cost, and the emission cost in the generalized cost function of the transportation system, respectively. In the remainder of this section, we present an analytical approximation of the components of the cost function (3).

3.1. Users cost

The generalized cost of traveling in the network basically depends on the mode of transport that users choose for their trips in the network. Assuming a flexibility of mode choice for users, a binomial logit model is employed to split the total travel demand of the system between the automobile and transit modes based on the average generalized cost (disutility)

of these modes for users. Approximating the share of travel demand of the automobile and transit modes, the total cost of the transportation system for network users (T_U) can be calculated as the summation of costs that automobile users (T_A) and transit riders (T_T) experience in the network:

$$T_U = T_A + T_T \quad (4)$$

In this paper, we make a use of the MFD model to capture the effect of congestion on the average speed and the outflow of the network in approximating the generalized cost that automobile users experience over the peak. The generalized cost of using the transit system for traveling in the network is also approximated on an aggregated level using an analytical model.

3.1.1. Modal split

In a bimodal transportation network, users evaluate the available modes of transport for their trips by comparing their characteristics, and users choose modes of transport that are aligned with their own preferences. The service quality that automobile and transit modes offer to users can be generally measured by the generalized cost (disutility) associated with using these modes for traveling in the network. Under the assumption that all network users have the option to choose between automobile and transit modes for their trips in the network, a binomial logit model can be employed to split the total demand of the transportation network between these modes² (Shabanpour et al., 2017a). Note that, in continuum approximation models, the focus is on aggregating the travel behavior of the users and ignoring fine details in order to answer big picture questions (Daganzo, 2010b; Daganzo et al., 2012; Nourinejad and Roorda, 2016). Accordingly, the share of travel demand of automobile and transit modes can be approximated based on the average generalized cost of using the automobile ($\tau_A^{avg} = T_A/N_A$) and transit modes ($\tau_T^{avg} = T_T/N_T$) for traveling in the network over the study period:

$$N_i = \frac{e^{-\gamma \tau_i^{avg}}}{\sum_{j \in \{A, T\}} e^{-\gamma \tau_j^{avg}}} N_{tot}, \quad i \in \{A, T\} \quad (5)$$

where N_i denotes the share of the total demand, N_{tot} , of automobile ($i = A$) and transit ($i = T$) modes, and γ is the constant parameter of the model that reflects the mode choice preferences of the users and can be estimated through calibration and reflects the aggregated preferences of users (Amirgholy et al., 2015).

In this paper, the total travel demand of the transportation system can have a general distribution over the study period. Assuming that the shape of the probability density function (PDF) of the temporal distribution of the demand, $f_w(t)$, is independent of the mode choices of users, the cumulative distribution of the wished schedules of automobile users for arriving to their destinations, $W(t)$, can be derived as below:

$$W(t) = N_A \int_0^t f_w(\omega) d\omega, \quad 0 \leq t \leq T_m \quad (6)$$

where T_m is the length of the study period. With the same line of reasoning, the temporal distribution of the transit demand can be derived to determine the capacity (prs/veh) required for the transit vehicles to fully meet the demand at all the stops across the service area over the peak period (see Appendix A).

3.1.2. Automobile users

In network peak periods where the vehicular accumulation of the system exceeds its optimal value, regional outflow drops below network capacity, and the average network speed dramatically decreases with the further rise in system accumulation. Consequently, users experience longer travel times in the network, and it becomes physically impossible for users to arrive at their destinations punctually. Under the perfect information assumption, rational users seek to minimize the generalized cost of their trips by adjusting their arrival times to a region. The cumulative result of the individual arrival time decisions leads to a user equilibrium condition in which no one can reduce his/her cost by changing his/her arrival time to the bottleneck. The general cost that users experience in the equilibrium condition of the network can be approximated by combining Vickrey's (1969) congestion theory with the MFD model.

3.1.2.1. Morning commute problem in urban networks. The morning commute problem is introduced in Vickrey's (1969) congestion theory to demonstrate the trip scheduling behavior of users in the equilibrium condition of a first-in, first-out single bottleneck with fixed capacity over the morning peak. Vickrey's (1969) model of dynamic congestion has been also elaborated in the literature by accounting for heterogeneity in commuter wished schedules, schedule deviation preferences, and value of time in modeling the equilibrium condition of the bottleneck (Henderson, 1974, 1977, 1981; Hendrickson and Kocur, 1981; Newell, 1987; Arnott et al., 1988; Arnott et al., 1992, 1994). The existence and uniqueness of the equilibrium solution of the morning commute problem also has been shown for a bottleneck with a general distribution of wished schedules (Smith, 1984; Daganzo, 1985) and a multiclass demand of commuters (Lindsey, 2004; Silva et al., 2014). Recent theoretical studies have also taken further steps towards generalizing the equilibrium and optimum solutions of the morning commute problem for continuously distributed preferences among users (Liu and Nie, 2011; Qian and Zhang, 2011; Van

² In designing multimodal transit systems, the binomial logit model used in this paper can be substituted by a nested logit model to split the travel demand of the region between different modes of transport.

den Berg and Verhoef, 2011a, 2011b; Xiao et al., 2011; Chen et al., 2015b; Liu et al., 2015; Wu and Huang, 2015; Amirgholy and Gonzales, 2017).

The equilibrium solution of the morning commute problem also has applications in optimizing transportation systems with time-dependent demand and state-variable capacity able to be modeled as queueing systems (Amirgholy and Gonzales, 2016). In recent studies, the congestion theory has been combined with the MFD model to solve the morning commute problem on an aggregated level for large urban regions with state-variable outflow (Small and Chu, 2003; Geroliminis and Levinson, 2009; Fosgerau and Small, 2013; Arnott, 2013; Fosgerau, 2015). As a result, the equilibrium conditions of the bottleneck can be extended to the network level by approximating the outflow of the network at each point in time as a function of the vehicular accumulation of the system having the NEF of the region as explained in Amirgholy and Gao (submitted for publication). The queueing diagram of Fig. 2a illustrates the equilibrium solution of the morning commute problem in an urban network.

In the queueing diagram of Fig. 2a, the cumulative distribution of the wished departure time of identical automobile users is represented by an S-shaped wished curve, $W(t)$, with a slope that exceeds the network capacity, μ_m , between two points in time and remains less outside. Consequently, the automobile users that arrive to the network (enter or start their trips inside the region) during the peak period (t_s to t_E) experience longer travel times in the network (N_Q users out of N_A users), while it becomes physically impossible for them to depart the network (exit or park inside the region) on time. So, we represent the actual departure of users from the network with the cumulative counts of vehicles departing the network by time t , $D(t)$, whose slope at each point in time equal to the outflow of the network at that point in time. The equilibrium arrival of the users to the bottleneck is also represented by the cumulative counts of the vehicles arriving to the network by time t , $A(t)$. Assuming that the trips of the identical automobile users have the same length in the network and the average speed of the network remains constant off the peak, the horizontal distance between the arrival and departure curves in the queueing diagram graphically represents the delay that users experience in their trips over the peak period. Similarly, the deviation of the users from their wished departure times can be also illustrated as the horizontal distance between the wished and actual departure curves, for early (departures between t_s and t_I) and late (departures between t_I and t_E) commutes. Under the perfect information assumption, rational users seek to adjust their arrival times to the network to minimize the generalized cost of their trips, which can be expressed as a linear combination of the off-peak travel time (τ_F), delay (τ_D), and earliness (τ_E) or lateness (τ_L) that they experience in their trips³:

$$\tau_A(N) = \tau_F + \tau_D(N) + e\tau_E(N) + l\tau_L(N) \quad (7)$$

where $\tau_A(N)$ denotes the general cost that user N experiences in his/her trip, and e and l are the earliness and lateness penalty factors that the identical travelers consider for deviating from their wished schedules in their trips, respectively. The cumulative result of individual arrival time decisions of the travelers leads to the user equilibrium condition in which no one can reduce his/her cost by unilaterally shifting his/her arrival time to the network. Following from Daganzo (1985), the equilibrium solution of the morning commute problem in urban networks can be generally described by two conditions.

In the equilibrium condition, the marginal cost of changing the arrival time to the network for each and every user equals zero. On this basis, the first condition relates the equilibrium arrival rate of the users (inflow of the network) at time t , $\dot{A}(t)$, to their departure rate (outflow of the network) at time $\Gamma = t + \tau_D(A(t))$, $\dot{D}(\Gamma)$, for the early and late departures over the peak period, t_s to t_E :

$$\dot{A}(t) = \begin{cases} \dot{D}(\Gamma)/(1 - e), & t \in [t_s, t_I - T_C] \quad \text{early users} \\ \dot{D}(\Gamma)/(1 + l), & t \in (t_I - T_C, t_E] \quad \text{late users} \end{cases} \quad (8)$$

while the outflow (departure rate) of the network at each point in time is determined by the NEF of the region based on the accumulation of the system (number of vehicles circulating inside the region) at that point in time, $n(\cdot)$:

$$\dot{D}(t) = \mu(n(\Gamma)) \quad (9)$$

In the morning commute problem, the accumulation of the system at each point in time can be approximated as the difference between the cumulative arrival and departure of users from the network at that point in time, which can be graphically represented as the vertical distance between the arrival and departure curves in the queueing diagram of Fig. 2a:

$$n(\Gamma) = A(\Gamma) - D(\Gamma) \quad (10)$$

The second equilibrium condition also determines the proportion of the lengths of the earliness period, $T_e = t_I - t_s$, to the lateness period, $T_l = t_E - t_I$, to be equal to the proportion of the lateness penalty factor of users to their earliness penalty factor:

$$\frac{T_e}{T_l} = \frac{l}{e} \quad (11)$$

³ It is also common in the literature to use α , β , and γ to express Eq. (7) equivalently as $\tau_A(N) = \alpha(\tau_F + \tau_D(N)) + \beta\tau_E(N) + \gamma\tau_L(N)$. In our notation, $e = \beta/\alpha$ and $l = \gamma/\alpha$.

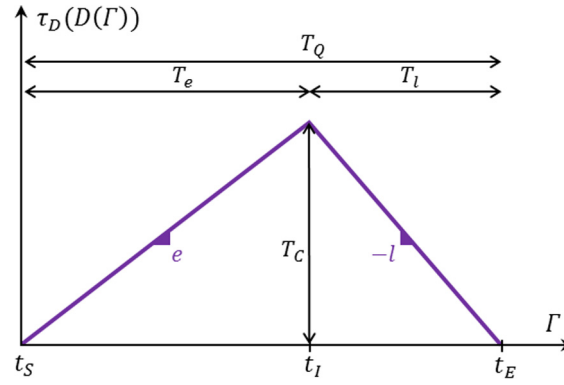


Fig. 3. Variation of delay with departure time of the travelers from the network.

network-wide design for the transit system. The exact, detailed methods can be then used to fine-tune the design of the transit system (Daganzo et al., 2012).

3.1.2.3. Generalized cost for automobile users. A transportation network has a certain capacity for serving the travel demand of users. In the peak period where the capacity of the network is insufficient to meet the demand, a rise in network vehicular accumulation beyond its optimal value would result in a drastic slowdown of flow in the network. Thus, automobile users experience longer travel times in the network, while they arrive to their destination earlier or later than they wish. So, the total cost of traveling in the network for the automobile users ($T_A = \sum_{N=1}^N \tau_A(N)$) can be expressed as the summation of the free-flow travel time (T_F), delay (T_D), earliness (T_E), and lateness (T_L) that they experience in the trips:

$$T_A = T_F + T_D + T_E + T_L \quad (16)$$

while, the total free flow time of automobile users, T_F , with average trip length L and free flow speed $v_F = \mu_m L / \bar{n}_m$ (see Fig. 1b) can be derived as below:

$$T_F = \frac{L}{v_F} N_A = \frac{\bar{n}_m}{\mu_m} N_A \quad (17)$$

The total delay that automobile users experience in the equilibrium condition can be graphically represented as the area between the arrival and departure curves in the queueing diagram of the network. Here, we make use of the bottleneck model to approximate the total delay of automobile users, T_D , as the triangular area between the arrival and departure curves of Fig. 2b:

$$T_D = \frac{el}{2\bar{\mu}(e+l)} N_Q^2 \quad (18)$$

Similarly, the total earliness and lateness costs that users experience in their trips also can be approximated as the area between the actual and wished departure curves weighted by the earliness (e) and the lateness (l) penalty factors that users consider for deviating from their schedules in their trips in the network. The area between the actual and wished departure curves can be estimated for general shapes of the wished curve using numerical integration techniques. However, in cases where the cumulative wished departure of users can be represented with an inverse Z-shaped curve with constant slope ω in the inclined part and zero elsewhere, as depicted in Fig. 4, the total earliness (T_E) and lateness (T_L) costs of traveling over the peak for automobile users can be analytically approximated as follows:

$$T_E = \frac{e}{2} \left(\frac{l}{e+l} \right)^2 \left(\frac{1}{\bar{\mu}} - \frac{1}{\omega} \right) N_Q^2 \quad (19)$$

$$T_L = \frac{l}{2} \left(\frac{e}{e+l} \right)^2 \left(\frac{1}{\bar{\mu}} - \frac{1}{\omega} \right) N_Q^2 \quad (20)$$

The total generalized cost that automobile users with an inverse Z-shaped wished curve experience in the network can be approximated as the summation of their free-flow travel time (T_F), delay (T_D), earliness (T_E), and lateness (T_L), according to the cost function (16).

It is also worth pointing out that delay and schedule deviation theoretically just occurs in the trips over the peaks ($\bar{n} > \bar{n}_m$) in which the capacity of the network is insufficient to meet the travel demand of the automobile users. Otherwise, the travel cost of automobile users in the network is limited to the free flow travel time of their trips when the network is not congested ($\bar{n} \leq \bar{n}_m$).

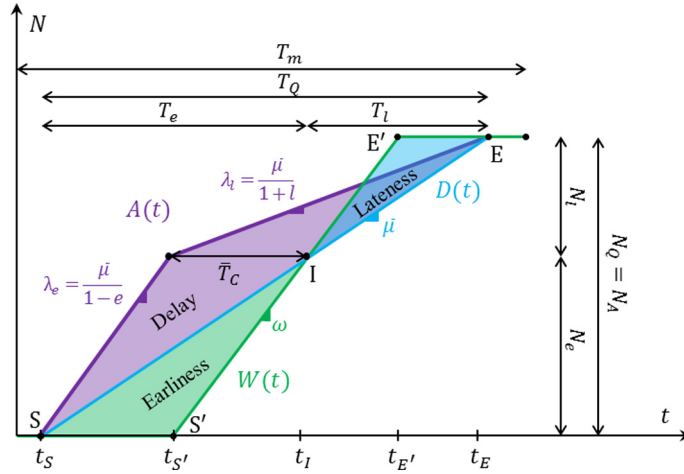


Fig. 4. Queuing diagram of the morning commute problem in urban network with an inverse Z-shaped wished curve.

3.1.3. Transit users

The generalized cost that transit riders experience in their trips directly depends on the spatiotemporal service quality that the transit system provides to the users. In a conventional transit system where the routes and schedules of vehicles are fixed, the total cost of using the service for the riders (T_T) can be expressed as a summation of the following: the sum total of their access (walking) time to stops, their waiting time at stops, and their waiting time at transfer points (T_W); the fare they pay to use the service (T_P); the time they spend in transit (T_I); and the penalty cost they consider for deviating from their wished schedules (T_{el}):

$$T_T = T_W + T_P + T_I + T_{el} \quad (21)$$

The components of the generalized cost of using the transit system for its users can be analytically approximated based on the network structure characteristics (stop spacing and line spacing) and operating characteristics (headway and fare) of the transit system. So, we formulate the problem for three prevalent transit system network allocation scenarios (mixed, dedicated, and parallel) in an urban region. In this section, we develop the model for a grid network structure as generally illustrated in Fig. 5; however, the proposed analytical model can be extended for other types of network structures like ring-radial as shown in Section 5. Here, the grid arterial network is expanded with fixed spacing D_N over an urban region with area A . The transit routes can also lay along the arterial network with line spacing D_T and stop spacing S , which remain fixed for the transit system all over the service region.

Assuming that demand is uniformly distributed over the region, the average access distance of the users to the closest stop at both the origin and destination sides of the trips is equal to $(S + D_T)/4$. Under the assumption that the users have no information regarding the schedule of the service prior to their trips, the average waiting time of the users for the transit system with the service headway H at the origin stops as well as the transfer points equals $H/2$. Assuming that transit users on average make δ transfers in their trips to get to their destinations, the total access and waiting cost of the users (T_W) can be approximated as below:

$$T_W = w \left(\frac{(1 + \vartheta\delta)H}{2} + \frac{S + D_T}{2v_w} \right) N_T \quad (22)$$

where, w denotes the penalty factor that users consider for access and waiting time in their trips, ϑ is the penalty factor for the time it takes the users to transfer between the transit lines, and v_w is their access (walking) speed.

The total fare that users pay for using the transit service (T_P) can be also expressed in units of time knowing the value of time of the users, VOT :

$$T_P = \left(\frac{p}{VOT} \right) N_T \quad (23)$$

where p denotes the fare of the transit system. Note that the agency may charge the heterogeneous users different fares based on the length of their trips in the network. However, the total fare that users pay to use the service can be still approximated based on the average fare charged by the agency using cost function (23).

The total in-transit time of users can also be approximated as a function of the average length of the trips using the transit system, L , as well as the commercial speed of the transit vehicles (miles traveled per unit of service time, including idling times at the stops), v_T :

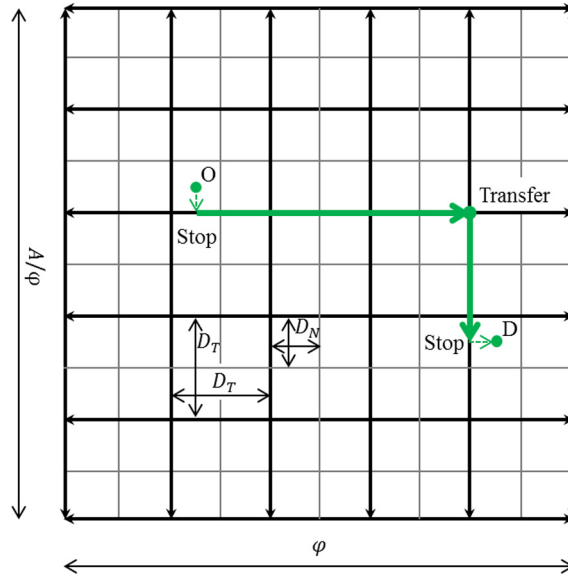


Fig. 5. Transit network structure (black lines) in a grid arterial network (gray lines).

$$T_I = \left(\frac{L}{\nu_T} \right) N_T \quad (24)$$

Note that the commercial speed of the transit vehicles varies with the characteristics of the system in different network allocation scenarios, as we explain later in Section 4.

Under condition that capacity of transit vehicles (prs/veh) is determined to meet the maximum demand rate over the peak (transit vehicles always have enough capacity (prs/veh) to pick up all the waiting users at all the stops over the service area during the peaks), and assuming that transit users have perfect information regarding the in-transit time of their trips (see Daganzo, 2010b), the average deviation occurring in the scheduling of their trips is equal to the summation of their average waiting time at the origin stop and at the transfer points, $(1 + \delta)H/2$. Since the probability of arriving early or late to the destinations is equal in each trip, the average weighted penalty factor for deviating from the wished schedules becomes equal to the average of their earliness and lateness penalty factors, $(e + l)/2$. As a result, the total cost of the schedule deviation for the transit users (T_{el}) can be approximated as below:

$$T_{el} = \left(\frac{e + l}{2} \right) \left(\frac{(1 + \delta)H}{2} \right) N_T \quad (25)$$

The total cost of using the transit system for users can be then approximated in units of time as the summation of the access and waiting time (T_w), equivalent cost of the fare (T_p), in-transit time (T_I), and schedule deviation cost (T_{el}), as represented by the generalized cost function (21). The shares of the automobile and transit modes from the total demand also can be determined using the binomial logit choice model (5) based on the average cost (disutility) of these modes for their users. The total cost that users experience in the network can be also expressed as the summation of the total cost associated with using automobile and transit modes in the network, as shown in (4).

3.2. Operating cost of transit system

The operating cost of the transit system (C_T) can be expressed as the summation of the costs associated with fleet size (C_M), the total vehicle miles traveled (C_{VMT}), the total vehicle hours traveled (C_{VHT}), the total revenue (negative cost) from the fare collection (C_R), constructing new infrastructure (C_C), and also the fixed cost (C_F) of operating the transit system:

$$C_T = C_M + C_{VMT} + C_{VHT} - C_R + C_C + C_F \quad (26)$$

The cost of acquiring and maintaining the transit vehicles constitutes a significant portion of the operating cost for the agency, which is directly proportional to the fleet size of the transit system. To approximate the fleet size of the system, we first derive the number of vehicles that the agency needs to provide the service headway of H in one of the lines with length ϕ , M_ϕ , by dividing the travel time of the line (ϕ/ν_T) by the headway of the service:

$$M_\phi = \frac{\phi}{\nu_T H} \quad (27)$$

So, the fleet size the agency needs to provide a service headway H in both directions in the grid network (horizontal and vertical lines) with line spacing D_T and service area A can be derived as below:

$$M = 2 \left(\frac{\varphi}{v_T H} \times \frac{A}{\varphi D_T} + \frac{A}{\varphi v_T H} \times \frac{\varphi}{D_T} \right) = \frac{4A}{v_T H D_T} \quad (28)$$

As a result, the total cost associated with the fleet size (C_M) can be approximated as follows:

$$C_M = \alpha_M \frac{4A}{v_T H D_T} \quad (29)$$

where α_M denotes the cost associated with one vehicle for the agency.

It follows that the total cost of the vehicle miles traveled (VMT) of the transit system (C_{VMT}) with fleet size M and commercial speed v_T over the study period T_m can be approximated as below:

$$C_{VMT} = \alpha_{VMT} M v_T T_m = \alpha_{VMT} \frac{4A}{H D_T} T_m \quad (30)$$

Similarly, the total cost of the vehicle hours traveled (VHT) of the transit system (C_{VHT}) over the study period T_m can be derived as follows:

$$C_{VHT} = \alpha_{VHT} M T_m = \alpha_{VHT} \frac{4A}{v_T H D_T} T_m \quad (31)$$

where α_{VMT} and α_{VHT} are the unit costs of VMT and VHT of the transit system for the agency, respectively.

The total revenue of the agency from collecting fares from riders (C_R) can be also calculated as below:

$$C_R = p N_T \quad (32)$$

Operation of the transit service may also require constructing new infrastructures along the lines in some of the network allocation scenarios (in particular, Metro). The cost associated with such construction (C_C) is generally proportional to the total length of the construction along the transit lines, $(\varphi + A/\varphi)/D_T$:

$$C_C = \alpha_C \frac{\varphi + \frac{A}{\varphi}}{D_T} \quad (33)$$

where α_C denotes the cost associated with constructing a unit of mile of the new infrastructure for the transit system.

3.3. Emission cost in urban network

The external cost of emissions in urban networks is another primary component of the generalized cost of the transportation system that should be taken into account in designing a sustainable transit system. Road traffic is the main source of air pollutants in urban regions, contributing up to 45%, 50%, and 90% of the total NO_x , HC, and CO emissions in the network, which can also rapidly rise with increasing congestion in the network (Olsson, 1994). There are different macro-scale emission models developed in the literature to approximate air pollutants emitted by motor vehicles as a function of the average speed of the network (Woensel et al., 2001; Sbayti et al., 2002; Nesamani et al., 2007; Wang et al., 2009; Shabanpour et al., 2017b). The emission of air pollutants in the network is also shown to be related to the total fuel consumption of vehicles in the region (Post et al., 1984; De Vlieger, 1997; Tong et al., 2000; Ahn et al., 2002; Rakha and Ding, 2003; Silva et al., 2009; Sobrino et al., 2016). In this research, we make a use of the MFD to approximate the average speed of the network in a VMT-based emission model, presented in Affum et al. (2003) to express the external cost of the emission (E_N) in the network as a function of the fuel consumption of the automobiles (F_A) and transit vehicles (F_T) in their total distance traveled in the network:

$$E_N = \alpha_E^A F_A + \alpha_E^T F_T \quad (34)$$

where the conversion factors α_E^A (\$/ml) and α_E^T (\$/ml) represent the cost of the air pollutants emitted by average automobiles and transit vehicles in the network, respectively, when consuming a volume unit of fuel.

The total fuel consumption of in-network automobiles, F_A , can be approximated for the total miles traveled by users with average trips L in the network, LN_A :

$$F_A = f_A L N_A \quad (35)$$

while, f_A (ml/km) denotes the fuel consumption of the average automobile per unit of distance traveled in the network, and can be approximated in urban networks (average speeds less than 50 km/h) based on Affum et al. (2003) as follows:

$$f_A = \frac{f_i^A}{v^{k_A}} + K_A \quad (36)$$

Here, f_i^A (ml/h) is the idling fuel consumption rate of the average automobile, and K_A and k_A are adjustment factors capturing characteristics of the vehicles and the network. The average speed of the network, v , is determined by the MFD model in different network allocation scenarios, as we explain in Section 4. As a result, the external cost of the emission in the region can be analytically approximated by substituting f_A from (36) into (35) and plugging the results back into the cost function (34).

The total fuel consumption of the transit system, F_T , can be similarly derived based on the total miles traveled of the transit system in the grid network over the study period T_m , $Mv_T T_m$, as below:

$$F_T = f_T M v_T T_m = \left(\frac{f_i^T}{(v_T)^{k_T}} + K_T \right) \frac{4A}{HD_T} T_m \quad (37)$$

where f_T (ml/km) is the fuel consumption of an average transit vehicle per unit of traveled distance, and f_i^T (ml/h) is its idling fuel consumption rate. K_T and k_T are also the adjustment factors of the model for the transit system. The commercial speed of the transit vehicles, v_T , is also derived for different network allocation scenarios in Section 4.

It is also worth pointing out that the amounts and types of air pollutants emitted by the automobile and transit modes largely depend on the technology used in these vehicles. In recent years, new technologies have significantly reduced the fuel consumption in the network by improving the efficiency of the motor power of the vehicles (McKain et al., 2000; Frey et al., 2007; Sharma et al., 2010; Oh et al., 2014; Lajunen, 2014; Alzuwayer et al., 2014; Khalilikhah et al., 2016).

4. Sustainable design of the transit system

A transit system with a competitive quality of service to automobiles can play a key role in alleviating the congestion and its adverse consequences for users and the environment in urban regions by attracting a considerable portion of the travel demand of the network over the peaks. However, the operating cost of the agency also escalates with improvement in the service quality of the transit system. In that sense, it is of great importance to make an optimal balance between the operating cost and service quality of the transit system by minimizing the generalized cost of the transportation system (C). To this end, we minimize a linear combination of the cost that users experience in the network (T_U), the operating cost of the transit system for the agency (C_T), and the external cost of the emission in the network (E_N), as expressed in cost function (3), by optimizing the line spacing (D_T), stop spacing (S), headway (H), and fare of the transit system (p) as follows:

$$\min_{D_T, S, H, p} C = \beta_T T_U + \beta_C C_T + \beta_E E_N \quad (38)$$

In this model, we split the total travel demand of the system (N_{tot}) between the automobile and the transit modes using logit model (5). To account for the interaction between automobile and transit modes in different network allocation scenarios, the MFD model is employed to approximate the commercial speed of the transit system (v_T) as well as the equivalent capacity ($\bar{\mu}$) of the bottleneck model and the average speed (v) of the arterial network in three different network allocation scenarios in urban regions: (i) mixed network (Bus), (ii) dedicated lanes (Bus Rapid Transit), and (iii) parallel networks (Metro). The optimization problem (38) falls into the category of a nonconvex Nonlinear Program (NLP), which can be solved using numerical methods (Chen et al., 2015a). In the numerical example of Section 6, we have solved the design problem using the GAMS/Baron optimization platform.

4.1. Scenario I: Mixed network (Bus)

In the first network allocation scenario, we consider the case where automobiles share the arterial network with a transit system with no dedicated lanes, e.g. a conventional bus system. In this case, the transit vehicles move with the traffic flow in the network, while the presence of the transit vehicles in the network also affects the dynamics of the congestion in the urban region. To account for the interaction between automobiles and transit vehicles in the mixed bimodal network, Geroliminis et al. (2014) proposes an exponential function for the three-dimensional MFD that relates the flow of the network ($Q(\bar{n}, M)$) to the accumulation of the automobiles (\bar{n}) and transit (M) vehicles as below:

$$Q(\bar{n}, M) = a_0 (\bar{n} + M) e^{a_1 \bar{n}^2 + a_2 M^2 + a_3 \bar{n}M + a_4 \bar{n} + a_5 M} \quad (39)$$

where parameters a_0 to a_5 should be estimated through calibration of the model using empirical data from the network so that the 3D-MFD model can provide an accurate approximation of the dynamics of the mixed congestion in bimodal networks. Alternatively, the effect of mixed traffic in the bimodal transportation network can be captured by converting the total number of transit vehicles into an equivalent number of automobiles, using the conversion factor θ , in calculating the accumulation of the system in the MFD model, although the accuracy of the model is less than the 3D-MFD. As a result, the equivalent capacity of the network can be approximated based on the relation (15) as below:

$$\bar{\mu} = \mu(n_{eq} + \theta M) \quad (40)$$

where n_{eq} can be derived using Eq. (14). On this basis, the average speed of the network over the peak period, v , can be derived by plugging $n = n_{eq} + \theta M$ into Eq. (2):

$$v = \frac{\mu(n_{eq} + \theta M)}{n_{eq} + \theta M + \bar{n}_m} L = \frac{\bar{\mu}}{n_{eq} + \theta M + \bar{n}_m} L \quad (41)$$

To derive the commercial speed of the transit system in the mixed urban network, we need to account for the loading and unloading time of the vehicles at the stops in the transit network, t_s . Assuming that t_s remains fixed for all the stops in the network over the peak, the commercial speed of the transit vehicles, v_T , can be approximated by dividing the distance it can travel without having stops in the network (which is equal to the average speed of the network, v) by the time it takes to travel the same distance in a network with stop spacing S (unit of time plus the total loading and unloading time of the vehicle per unit of time, $1 + t_s v_T/S$):

$$v_T = \frac{v}{\left(1 + \frac{t_s v_T}{S}\right)} \quad (42)$$

So, the commercial speed of the transit vehicles in the mixed network, v_T , can be determined by solving Eq. (42) for $v_T \geq 0$ as below:

$$v_T = \frac{\sqrt{S(S + 4t_s v)} - S}{2t_s} \quad (43)$$

Approximating the equivalent capacity of the network ($\bar{\mu}$), average network speed (v), and commercial speed of the transit system (v_T) using Eqs. (40), (41), and (43), the total cost of the mixed bimodal urban network can be derived as explained in Section 3. The network structure (line spacing and stop spacing) and operating characteristics (headway and fare) of the transit system can then be optimized by minimizing the total cost of the network, as presented in mathematical problem (38).

4.2. Scenario II: Dedicated lanes (Bus Rapid Transit)

In the second scenario, we allocate a portion of the capacity of the arterial network to the transit system by dedicating a number of lanes to the transit vehicles, e.g. via a Bus Rapid Transit (BRT), in order to improve the mobility of users in the urban region. Assuming that the dedicated lanes are uniformly distributed over the region, the exit function of the new network can be derived by rescaling exit function (1) for the original network, as explained in Gonzales and Daganzo (2012):

$$\mu(n) = \frac{\rho A_N}{L} Q\left(\frac{n + \rho \bar{n}_m}{\rho A_N}\right) \quad (44)$$

where the network allocation factor ρ can be calculated as the ratio of the network area used by the automobiles, after dedicating a portion of the area to the transit system (A_T), to the total area of the original network (A_N):

$$\rho = \frac{A_N - A_T}{A_N} \quad (45)$$

Here, A_T and A_N can be approximated in terms of the widths of the transit lanes, d_T , and the arterials, d_N , in a two-way grid network as below:

$$A_T = \frac{2d_T\left(\varphi + \frac{A}{\varphi}\right)}{D_T} \quad (46)$$

$$A_N = \frac{2d_N\left(\varphi + \frac{A}{\varphi}\right)}{D_N} \quad (47)$$

By plugging A_T and A_N from Eqs. (46) and (47) into (45), and with a little manipulation, relationship (45) can be rewritten as below:

$$\rho = 1 - \frac{D_N d_T}{D_T d_N} \quad (48)$$

As a result, the equivalent capacity of the network, $\bar{\mu}$, can be approximated by plugging n_{eq} from Eq. (14) into the new network exit function (44):

$$\bar{\mu} = \mu(n_{eq}) = \frac{\rho A_N}{L} Q\left(\frac{n_{eq} + \rho \bar{n}_m}{\rho A_N}\right) \quad (49)$$

On this basis, the average speed of the automobiles in the network, v , can be also derived according to relation (2):

$$v = \frac{\bar{\mu}}{n_{eq} + \rho \bar{n}_m} L \quad (50)$$

In the second network allocation scenario, where the transit system uses dedicated lanes to carry passengers in the network, the commercial speed of the vehicles is independent of the traffic condition of the arterial network. In this case, the travel time of the transit system between two consecutive stops, τ_s , can be derived using the kinematic equation for constant acceleration motion by decomposing the stop spacing S into accelerating and decelerating parts with the same length, $S/2 = \frac{1}{2}a_T(\tau_s/2)^2$:

$$\tau_s = 2\sqrt{\frac{S}{a_T}} \quad (51)$$

where a_T denotes the absolute value of the maximum allowable (tolerable) acceleration/decelerate rate of the transit vehicles. Assuming that the implemented transit signal priority (TSP) strategy in the arterial network can reduce the delay of the transit system at intersections close to zero (Christofa and Skabardonis, 2011; Christofa et al., 2013), the commercial speed of the transit system, v_T , can be approximated by dividing the distance between two consecutive stops, S , by the time it takes to load/unload the passengers at the first stop and carry them to the next stop, $t_s + \tau_s$:

$$v_T = \frac{S}{t_s + 2\sqrt{\frac{S}{a_T}}} \quad (52)$$

Having the equivalent capacity of the network ($\bar{\mu}$), average speed of the automobiles (v), and commercial speed of the transit system (v_T) using Eqs. (49), (50), and (52), the network structure (line spacing and stop spacing) and operating characteristics (headway and fare) of the transit system with dedicated lanes can be optimized by minimizing the total cost of the transportation system in the objective function of problem (38).

4.3. Scenario III: Parallel network (Metro)

In the third network allocation scenario, the transit system operates in a separate network parallel to the arterial network, e.g. a Metro. In this case, it becomes possible to improve the mobility of users in the network without reducing the capacity of the arterial network. Hence, Scenario III can be generally viewed as a special case of Scenario II in which the network allocation factor ρ is kept equal to unity at the cost of constructing new infrastructure for the transit system in a parallel network. So, operation of the transit system has no effect on the traffic flow of automobiles, and the outflow of the network can be approximated by setting $\rho = 1$ in the network exit function (49). The average speed of the arterial network, v , can be similarly derived using Eq. (50). The commercial speed of the transit system, v_T , can also be calculated using Eq. (52) due to the similarities in operation of the transit systems in scenarios II and III, while the allowable acceleration/deceleration rate of the transit vehicles, a_T , can be higher when the transit vehicles are operating in a separate network. The network structure (line spacing and stop spacing) and operating characteristics (headway and fare) of the transit system in a parallel network similarly can be optimized by minimizing the total cost of the transportation system in the objective function of problem (38).

5. Model extension for ring-and-radial network structure

In Sections 3 and 4, we have presented an optimal design for sustainable transit systems in grid urban networks. However, the proposed model can be also extended for other types of network structures by reformulating the components of the total cost according to the geometry of the network. In this section, we extend the proposed continuum approximation model for a ring-and-radial network structure using a polar coordinate system. To this end, we reformulate the components of the total cost needed to be adapted for a ring-and-radial transit network (with radial ring-line spacing r_T and angular radial-line spacing θ_T) in a circular service area with radius R , as illustrated in Fig. 6.

In this case, the generalized cost that automobile users experience in the arterial network (with radial ring-arterial spacing r_N and angular radial-arterial spacing θ_N) (T_A) can be approximated using the macroscopic model developed by combining congestion theory with the MFD in Section 3.1.2. The components of the generalized cost associated with using transit system for traveling in the ring-and-radial network (T_T) can be also approximated by adopting the analytical model presented in Section 3.1.3. In generalized cost function (21), the access and waiting time for the users (T_w) is the component of the cost that varies with change in the geometry of the network, so it should be adapted for the ring-and-radial network structure while the fare that transit riders pay to use the service (T_p), the time they spend in-vehicle (T_i), and the cost associated with deviations in their wished schedules (T_{el}) can be approximated using cost functions (23)–(25) in the network with ring-and-radial structure as well. In ring-and-radial networks, the average access distance of the users to the closest stop on the ring-lines equals to $(r_T + S)/4$, while this distance for the stops on the radial-lines equals to $(R\theta_T/2 + S)/4$. With this change in the average access distance of the users to the stops, the total access and waiting time of the transit users in the ring-and-radial network can be derived by reformulating the cost function (22) as follows:

$$T_w = w \left(\frac{(1 + \vartheta\delta)H}{2} + \frac{S + \frac{1}{2}(r_T + \frac{R\theta_T}{2})}{2v_w} \right) N_T \quad (53)$$

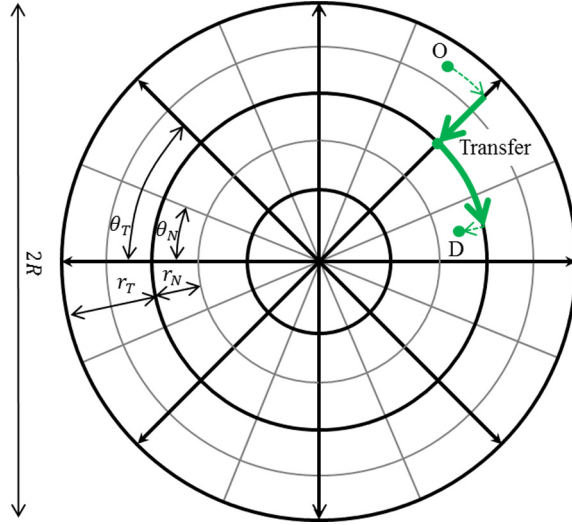


Fig. 6. Transit network structure (black lines) in a ring-and-radial arterial network (gray lines).

The analytical model developed in Section 3.2 to approximate the operating cost of the transit system for the agency can be also adapted by recalculating the required fleet size to run the service in the ring-and-radial network. In this case, the fleet size required to provide service headway H in each of the radial-links (with length R), M_r , and in the average ring-link (with length πR), M_θ , can be derived based on Eq. (27) as below:

$$M_r = \frac{R}{V_T H} \quad (54)$$

$$M_\theta = \frac{\pi R}{V_T H} \quad (55)$$

Following from above, the fleet size that agency needs to provide service headway H in both directions of the ring-and-radial network can be approximated by reformulating relation (28) as below:

$$M = 2 \left(\frac{R}{V_T H} \times \frac{2\pi}{\theta_T} + \frac{\pi R}{V_T H} \times \frac{R}{r_T} \right) = \frac{2\pi R}{V_T H} \left(\frac{2}{\theta_T} + \frac{R}{r_T} \right) \quad (56)$$

Accordingly, the costs associated with fleet size (C_M), VMT (C_{VMT}), and VHT (C_{VHT}) of the transit system can be derived for the ring-and-radial network as follows:

$$C_M = \alpha_M \frac{2\pi R}{V_T H} \left(\frac{2}{\theta_T} + \frac{R}{r_T} \right) \quad (57)$$

$$C_{VMT} = \alpha_{VMT} M v_T T_m = \alpha_{VMT} \frac{2\pi R}{H} \left(\frac{2}{\theta_T} + \frac{R}{r_T} \right) T_m \quad (58)$$

$$C_{VHT} = \alpha_{VHT} M T_m = \alpha_{VHT} \frac{2\pi R}{V_T H} \left(\frac{2}{\theta_T} + \frac{R}{r_T} \right) T_m \quad (59)$$

The cost of constructing new infrastructures along the transit lines (C_C) in the ring-and-radial network can be also approximated based on cost function (33) as below:

$$C_C = \alpha_C \left(\frac{2\pi}{\theta_T} \times R + \pi R \times \frac{R}{r_T} \right) = \alpha_C \left(\frac{2\pi R}{\theta_T} + \frac{\pi R^2}{r_T} \right) \quad (60)$$

Having the components of the cost, the operating and construction cost of the transit system (C_T) can be approximated using the generalized cost function (26), in which the revenue of the agency from fare collection (C_R) can be calculated using relation (32).

The VMT-based model presented in Section 3.3 to approximate the emission cost of transportation system can be also extended for the ring-and-radial network structure by recalculating the total fuel consumption of the transit system (F_T) based on the VMT of the transit system in the ring-and-radial network:

$$F_T = f_T M v_T T_m = \left(\frac{f_i^T}{(v_T)^{k_T}} + K_T \right) \left(\frac{2\pi R}{H} \left(\frac{2}{\theta_T} + \frac{R}{r_T} \right) \right) T_m \quad (61)$$

The external cost of emission (E_N) in the ring-and-radial network can be then approximated using cost function (34) in which the total fuel consumption of in-network automobiles (F_A) can be calculated using relations (35) and (36) as explained in Section 3.3.

The design of the sustainable transit system can be then optimized by minimizing a linear combination of the of the cost that users experience in the network (T_U), the operating cost of the transit system for the agency (C_T), and the external cost of the emission in the network (E_N), as explained in Section 4:

$$\min_{r_T, \theta_T, S, H, p} C = \beta_T T_U + \beta_C C_T + \beta_E E_N \quad (62)$$

where radial ring-line spacing (r_T), angular radial-line spacing (θ_T), stop spacing (S), headway (H), and fare of the transit system (p) are the decision variables of the design problem in different network allocation scenarios. In the first network allocation scenario, the fleet size of the agency in relations (39)–(41) should be approximated using Eq. (56). In the second scenario, the network allocation factor (ρ) needs to be recalculated based on the dedicated area to the transit system (A_T) and the total area of the original network (A_N) in the ring-and-radial network:

$$A_T = 2d_T \left(\frac{2\pi R}{\theta_T} + \frac{\pi R^2}{r_T} \right) \quad (63)$$

$$A_N = 2d_N \left(\frac{2\pi R}{\theta_N} + \frac{\pi R^2}{r_N} \right) \quad (64)$$

The third network allocation scenario, however, can be viewed as a special case of Scenario II in which the network allocation factor ρ is kept equal to unity with no need to further adaption.

The sustainable design transit system in a ring-and-radial network can be optimized by solving problem (62) as explained in Section 4.3.

6. Numerical example

In this section, we provide a numerical illustration of the proposed continuum approximation model to compare the optimal design of the transit system in different network allocation scenarios. The transit system is designed to provide a comparable level of service to the automobile for users in a grid arterial network with spacing $D_N = 0.5$ km and free flow speed of $v_f = 40$ km/h in a square urban region (see Fig. 5) in which the total travel demand of the transportation system is uniformly distributed over an area of $A = 100$ km² ($\varphi = 10$ km). The wished schedule of users is also assumed to be uniformly distributed over the period of 1 h ($f_w(t) = 1$, $t_s \leq t \leq t_E$, and zero otherwise), which can be graphically represented by an inverse Z-shaped wished curve with slope of $\omega = N_A = N_Q$ in the inclined part (see Fig. 4). The effect of congestion in the homogeneous urban network over the peak period is also captured using the declining part of the NEF of the region ($\bar{n}_m = 3000$ veh), which is assumed to be properly estimated by a quadratic function as below:

$$\mu(n) = \frac{\mu_m}{n_g^2} (n_g^2 - n^2) \quad (65)$$

where $\mu_m = 25,000$ veh/h and $n_g = 11,000$ veh. The total travel demand of identical users with a trip length of $L = 4.8$ km splits between the automobile and transit modes using a binomial logit model with constant parameter $\gamma = 1.20$. The total cost that automobile users with schedule penalty factors of $e = 0.5$ and $l = 1$ experience in their morning trips over the study period (assumed to be equal to the peak period in this example, $T_p = N_A/\bar{\mu}$) is approximated using the bottleneck model with an equivalent capacity for the network. The total cost of using the transit service can be also approximated using the proposed analytical model, with a loading and unloading time of $t_s = 1$ min, in trips with a $\delta = 1$ transfer on average per trip, for users with a walking speed of $v_w = 2$ km/h, a waiting penalty factor of $w = 1$, and a value of time of $VOT = 10$ \$/h. The operating cost of the transit system is also approximated for the values of the cost coefficients and the system parameters that are listed in Table 1 for different network allocation scenarios. The cost coefficients and the model parameters for the emission cost of the automobiles and transit system in the network are also summarized in Table 2. In this example, the network structure and the operating characteristics of the sustainable transit system can be designed by minimizing the total cost of the transportation network:

$$C = 10T_U + C_T + 3E_N \quad (66)$$

Mathematical program (38) is a non-convex optimization problem which can be numerically solved using the GAMS/BARON platform (Sahinidis, 1996; Tawarmalani and Sahinidis, 2004). To solve the numerical example here, we set the optimality gap for Baron to 1% and the time limit for the solver to 5000 sec. With such presetting, the average and maximum computational times for solving the numerical example are 5.6 sec and 60.43 sec, respectively. Table 3 compares the optimal

Table 1

Operating cost coefficients and system parameters of the transit service in Scenarios I, II, and III.

Scenario	Operating cost coefficients					System parameters		
	α_M (\$/veh)	α_{VMT} (\$/km)	α_{VHT} (\$/h)	α_C (\$/km)	C_F (\$)	θ	d_T/d_N	a_T (m/s ²)
I	700	0.7	1.4	0	70,000	2	N/A	N/A
II	700	0.7	1.4	0	70,000	N/A	1/2	1.0
III	7000	0.7	1.4	70,000	70,000	N/A	N/A	1.5

Table 2

Emission cost coefficients and fuel consumption model parameters of the transportation system in Scenarios I, II, and III.

Scenario	Emission cost coefficients		Fuel consumption parameters					
	Automobile	Transit	Automobile			Transit		
	α_E^A (\$/ml)	α_E^T (\$/ml)	f_i^A (ml/h)	K_A (ml/km)	k_A	f_i^T (ml/h)	K_A (ml/km)	k_A
I, II	5×10^{-4}	8×10^{-4}	1000	50	1	3000	200	1
III	5×10^{-4}	N/A	1000	50	1	N/A	N/A	N/A

Table 3

Optimal design of the transit service and minimized costs of the transportation system in Scenarios I, II, and III.

N_{tot} (prs)	Scenario	Transit design variables					Modal speeds		Modal splits		Average users' costs	
		D_T (km)	S (km)	H (h)	p (\$)	M (veh)	v (km/h)	v_T (km/h)	N_A (prs)	N_T (prs)	τ_A^{avg} (h)	τ_T^{avg} (h)
45,000	No transit	N/A	N/A	N/A	N/A	N/A	6.36	N/A	45,000	N/A	1.40	N/A
	I	2.20	0.74	0.27	3.55	81.89	9.92	8.36	33,625	11,375	0.57	1.47
	II	1.16	1.32	0.14	1.58	69.28	10.62	35.84	25,610	19,390	0.54	0.77
	III	4.20	2.05	0.29	4.17	6.05	9.79	55.13	34,936	10,064	0.60	1.60
50,000	No transit	N/A	N/A	N/A	N/A	N/A	4.69	N/A	50,000	N/A	1.68	N/A
	I	2.02	0.72	0.24	3.16	109.00	8.77	7.48	36,236	13,764	0.67	1.47
	II	1.08	1.28	0.13	1.40	80.54	9.67	35.05	26,951	23,049	0.61	0.74
	III	3.43	1.89	0.25	3.00	8.93	9.49	51.89	35,698	14,302	0.63	1.39
55,000	No transit	N/A	N/A	N/A	N/A	N/A	3.13	N/A	55,000	N/A	2.48	N/A
	I	1.90	0.71	0.23	2.87	138.59	7.69	6.65	38,802	16,198	0.78	1.51
	II	1.03	1.25	0.12	1.25	91.08	8.88	34.41	28,090	26,910	0.68	0.72
	III	3.01	1.79	0.23	2.47	11.50	8.98	49.92	36,992	18,008	0.67	1.27
60,000	No transit	N/A	N/A	N/A	N/A	N/A	1.67	N/A	60,000	N/A	4.20	N/A
	I	1.63	0.60	0.23	2.83	194.30	6.29	5.46	42,141	17,859	0.98	1.84
	II	0.99	1.23	0.12	1.13	100.77	8.21	33.89	29,050	30,950	0.75	0.70
	III	2.73	1.72	0.22	2.11	13.91	8.49	48.49	38,268	21,732	0.72	1.19

design and performance of the transportation system without and with a transit system in different network allocation scenarios for different levels of travel demand in the network. As can be inferred from comparing the average speed of the automobiles (v) in the network in absence of a transit system in Table 3, a rise in the travel demand of the network (N_{tot}) can significantly affect the performance of the transportation system by moving the network towards complete gridlock. However, a competitive transit system can alleviate the drop in average speed of the network by attracting a considerable portion of the travel demand in the network. Table 3 summarizes the change in the optimal design of the transit system and the performance of the transportation system in different network allocation scenarios with a rise in the total demand of the system. By comparing the primary design variables of the system in Table 3, it can be observed that the optimal line spacing (D_T), stop spacing (S), headway (H), and fare (p) of the transit system generally decrease (improve) with a rise in the total demand of the system in all network allocation scenarios, which naturally gives a boost to the need for a larger fleet size (M) to meet the optimal quality of service. The improvement in the service quality of the optimized transit system with the rise in demand is directly reflected in its share of the total demand of the system. Fig. 7 plots the rise in the demand share of the transit system in different network allocation scenarios with the increase in the travel demand of the region. In this example, the results show that the transit system with dedicated lanes (Scenario II) can attract a larger portion of the travel demand in comparison to other network allocation scenarios by offering a better quality of service to the users. The results also show that the share of the transit system in Scenario III is higher than Scenario I only when the total demand of the system is high enough.

To optimize the design of the transits system in different scenarios, we minimize a linear combination of the social, economic, and environmental costs of the transportation system, which generally escalate with a rise in the total demand of the system, as illustrated in Fig. 8a–d. In this example, the total cost of the transportation system (C) in Scenario II turns out to be

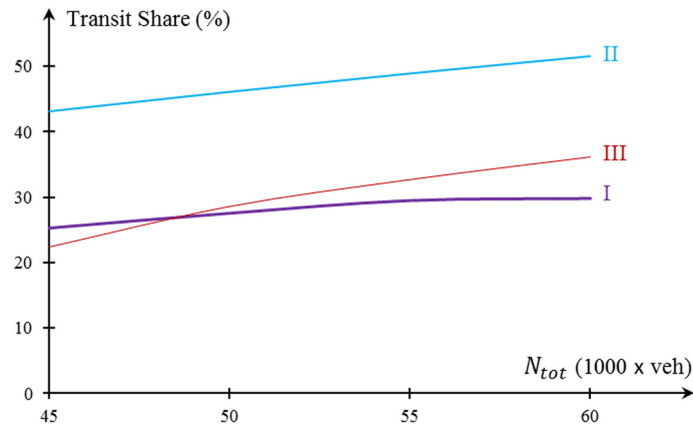


Fig. 7. Variations in modal share of the transit system with change in total travel demand of the region in different network allocation scenarios.

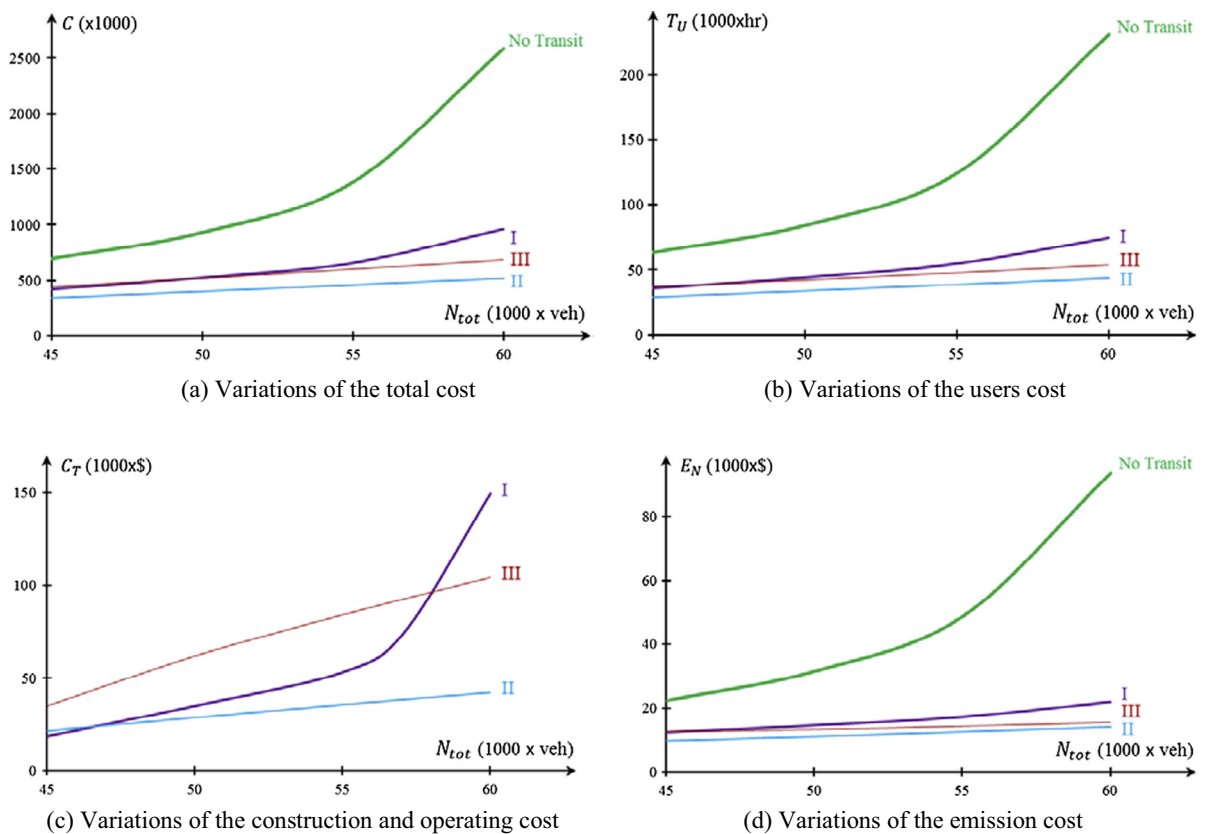


Fig. 8. Variations of the costs of the transportation system with change in the total demand of the system in different scenarios.

lower than in scenarios I and III, while a rise in demand can barely affect the escalation rate of the cost (slope of the cost curve), as illustrated in Fig. 8a. It is also worth pointing out that the transit system with a separate network (Scenario III) can reduce the total cost more than the mixed network (Scenario I) only when the total demand of the system is high enough; otherwise, the total cost of the system drops below that in the mixed network (Scenario I). The difference in the total cost of the transportation system in different network allocation scenarios can be attributed to the difference in the values of the cost components in those allocation scenarios.

Fig. 8b plots the variations of the total generalized cost that users experience in their trips (T_U) with a rise in the total travel demand of the system in different network allocation scenarios. In this example, the generalized cost of traveling in the network in Scenario II is lower than that for scenarios II and III. In spite of the higher construction and operating costs

of the transit system (C_T) in parallel networks, the generalized cost that users experience in Scenario III is not even lower than the cost for Scenario I when the demand of the system is not high enough. The optimal line spacing, stop spacing, and fare of the transit system in Scenario III sharply increase with the decline in the travel demand of the region due to the high construction and operating cost of the transit system in parallel networks. The consequent rise in the access and waiting cost (T_W) and the fare (p) of the transit system in Scenario III significantly affects the quality of the service such that it cannot even be compensated by the higher commercial speed of the transit vehicles (v_T) in parallel networks. This result also coincides with the general guidelines for designing express transit systems, which suggests providing a high-quality feeder system for the express service in the region in order to improve the efficiency of the transit system by reducing the access time of the system (Daganzo, 2010b). The variations in the construction and operating cost of the transit system with a rise in demand in different network allocation scenarios is also depicted in Fig. 8c. In this example, the construction and operating cost of the transit system in Scenario II is lower than that of Scenario III, in which high-speed vehicles operate in a parallel network. Interestingly, the operating cost of the transit system in Scenario I dramatically increases with the rise in the total demand of the system from values below Scenario II to values above Scenario III due to the sharp decline in the commercial speed of the transit vehicles (V_T) in the congested mixed network. In this example, the emission cost of the transportation system (E_N) in Scenario II also turns out to be lower than that for Scenario I, and even lower than the emission cost for Scenario III, as illustrated in Fig. 8d. In general, it can be concluded from this example that an optimally designed transit system can significantly reduce the total cost of the transportation network over the peaks. In the other words, the socio-environmental gain from operating an optimized transit system (in all network allocation scenarios) is higher than the associated operating cost. However, the efficiency of the transit system in different network allocation scenarios largely depends on characteristics of the demand over the peak in the network.

7. Conclusion

Transit systems can play a key role in improving user mobility and reducing emissions in urban networks. Urban network congestion can be reduced by attracting a considerable portion of the total demand to the transit system. However, the share taken by the transit system from the total travel demand of the network largely depends on providing users with a quality of service that is competitive to automobile use. Yet, improving the service quality of the transit system presents the challenge of also escalating the operating costs of the system. The objective of this research is to strike a balance between the service quality and operating cost of the transit system by minimizing the total cost of the transportation network. To this end, we propose a continuum approximation model that optimizes the network structure (line spacing and stop spacing) and the operating characteristics of the transit system (headway and fare) by minimizing a linear combination of the generalized cost that users experience in their trips, the operating cost of the transit system, and the external cost of emissions in an urban network. The model is first developed for a grid network structure, and then extended for a ring-and-radial network. In this research, we make use of a binomial logit choice model to split the total demand of the transportation system between the automobile and transit modes based on the disutility of these modes for the users. To account for the effect of congestion on the generalized cost that automobile users experience over the morning peak in the network, Vickrey's (1969) congestion theory is combined with the MFD to formulate the network problem as a bottleneck model. We have also developed an analytical model for the transit system in which the effect of the congestion on the service quality and performance of the transit system is taken into account using the MFD model. This analytical model is used to approximate the generalized cost that transit users experience in their trips as well as the operating and construction cost of the transit system for the agency. To account for the effect of congestion on the external cost of air pollutants emitted by automobiles and transit vehicles in the network, we extend a VMT-based fuel consumption model by approximating the average speed of the network using the MFD.

In the transit design problem, we approximate the average speed and the equivalent capacity of the arterial network as well as the commercial speed of the transit system in three primary network allocation scenarios in urban regions. In the first scenario, we consider the case where automobiles share the arterial network with transit vehicles with no dedicated lanes (Bus transit). In the second scenario, however, we allocate a portion of the capacity of the arterial network to the transit system by dedicating number of lanes to the transit vehicles in order to improve the mobility of users in the urban region (Bus Rapid Transit, BRT). In the third scenario, the transit system operates in a separate network parallel to the arterial network in order to improve user mobility without reducing the capacity of the arterial network (Metro). Efficiency of each of the network allocation scenarios in reducing the total cost of the system largely depends on the level of travel demand and trip length of the users in the region. The continuum approximation model proposed in this paper can be extended for designing multimodal transit systems in large urban networks with multiclass users by employing a nested logit model for the modal split of the users.

While this paper focuses on formulating the sustainable transit design problem in urban networks with grid and ring-and-radial structures, the model can be extended as well for other types of network structures, like hybrid and hub-and-spoke. Moreover, the structure of the transit network can be further elaborated by considering hierarchies in designing the transit system, i.e., express and local services. The key contribution of the research is to develop a continuum approximation model that can be used to obtain the primary design guidelines of the sustainable transit systems in congested urban

networks. The design guidelines of the transit system can be translated into a practicable elaborate plan in the fine-tuning step using detailed numerical techniques.

Acknowledgments

This work was supported in part by National Science Foundation project CMMI-1462289 and the Natural Science Foundation of China (NSFC) project # 71428001.

Appendix A. Capacity of transit vehicles

The primary design guidelines of the transit system are derived in this paper assuming that transit vehicles have enough capacity (prs/veh) to serve the demand in all the stops across the service area over the peak period. The vehicle capacity required to serve the peak demand can be then determined based on the maximum demand rate for the transit service in the region. In this problem, the temporal distribution of the transit demand over the study period ($q_T(t)$) can be obtained by scaling the probability density function (PDF) of the temporal distribution of the demand ($f_w(t)$) as below:

$$q_T(t) = N_T f_w(t), \quad 0 \leq t \leq T_m \quad (\text{A.1})$$

where N_T denotes the share taken by the transit system from the total travel demand of the region (see Section 3.1.1), and T_m is the length of the study period. Fig. A.1 illustrates a hypothetical distribution of transit demand with the maximum rate q_T^m over the study period $0 \leq t \leq T_m$.

Under the assumption that the transit demand is uniformly distributed across the service area A , the maximum accumulation of demand at each of the stops in any of 4 directions of a grid network (with line spacing D_T and stop spacing S as illustrated in Fig. A.2) in a time gap between arrivals of two consecutive transit vehicles with headway H can be derived as below:

$$Q_T^1 = H(D_T S) \left(\frac{q_T^m}{4A} \right) \quad (\text{A.2})$$

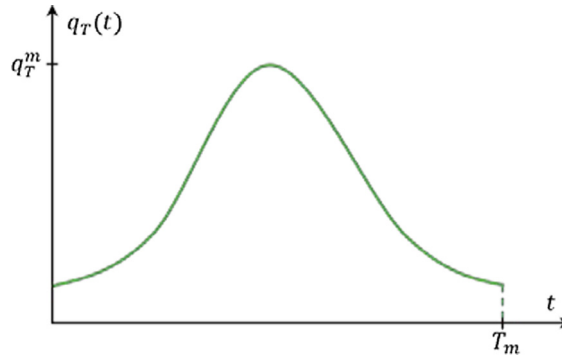


Fig. A.1. Hypothetical distribution of the transit demand over the peak.

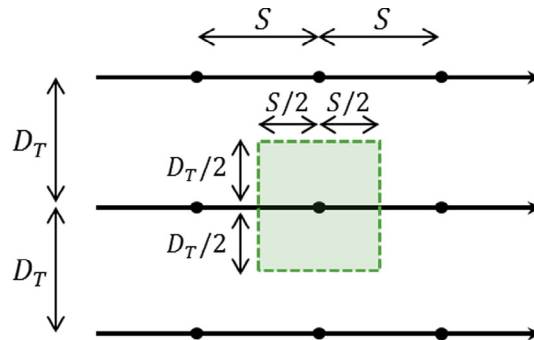


Fig. A.2. Demand attraction area of a single stop in a grid network.

Following from above, the capacity required for the transit vehicles (Q_T^m) to meet the demand at each and every stops in the service area over the peak period can be determined by multiplying the maximum accumulation of the transit riders with average trip length L at each stop and the number of stops the transit riders stay on board, $\left[\frac{L}{S}\right]$:

$$Q_T^m = Q_T^1 \left[\frac{L}{S} \right] = \left(\frac{HD_T S q_T^m}{4A} \right) \left[\frac{L}{S} \right] \quad (\text{A.3})$$

Appendix B. Glossary of symbols

See Table 4.

Table 4
Variable description.

Symbol	Description
a_T	absolute value of the maximum allowable acceleration/decelerate rate of the transit vehicles
A	region area
$A(t)$	cumulative counts of vehicles arriving to the network by time t
A_N	total area of the network
A_T	total area of the transit network
C	transportation system total cost
C_C	cost associated with constructing new infrastructures
C_F	fixed operating cost of transit system
C_M	cost associated with fleet size
C_T	transit system operating cost
C_{VMT}	cost associated with total vehicle miles traveled
C_{VHT}	cost associated with total vehicle hours traveled
C_{VHT}	total revenue from fare collection
d_N	average width of the arterials
d_T	width of the transit lanes
D_T	transit system line spacing in the grid network
D_N	arterial spacing in the grid network
$D(t)$	cumulative counts of vehicles departing the network by time t
e	earliness penalty factor that the identical travelers consider for deviating from their wished schedules
E_N	transportation system emission cost
f_A	fuel consumption of the average automobile per unit of distance traveled in the network
f_i^A	idling fuel consumption rate of the average automobile
F_A	total fuel consumption of in-network automobiles
$f_w(t)$	probability density function (PDF) of the temporal distribution of the demand
f_T	fuel consumption of the average transit vehicle per unit of distance traveled in the network
f_i^T	idling fuel consumption rate of the average transit vehicle
F_T	total fuel consumption of transit vehicles
H	transit system service headway
k	network vehicular density
K_A, k_A	adjustment factors capturing characteristics of the automobiles and the network
K_T, k_T	adjustment factors capturing characteristics of the transit vehicles and the network
l	lateness penalty factor that the identical travelers consider for deviating from their wished schedules
L	average trip length of the users in the network
n	vehicular accumulation of the system, $n = \bar{n} - \bar{n}_m$
\bar{n}	number of the vehicles circulating inside the region
n_{eq}	equivalent accumulation of the network in the bottleneck model
\bar{n}_m	optimum number of the vehicles circulating inside the region
M	fleet size
M_r	number of vehicles needed in the radial-links (with length R) in the ring-and-radial network
M_θ	number of vehicles needed in the average ring-link (with length πR) in the ring-and-radial network
M_ϕ	number of vehicles needed in the lines (with length ϕ) in the grid network
N_{tot}	total travel demand of the region
N_A	automobile travel demand
N_T	transit travel demand
N_Q	peak travel demand of automobile
p	transit fare
q	network flow
q_T^m	maximum rate for transit service over the peak period
Q_T^m	capacity required for the transit vehicles
$q_T(t)$	temporal distribution of the transit demand over the study period
$Q(k)$	MFD of the network
r_D	radial ring-line spacing

(continued on next page)

Table 4 (continued)

Symbol	Description
r_N	radial ring-arterial spacing
R	radius of the region in the ring-and-radial network
S	transit system stop spacing
t	time
t_s	loading and unloading time of the transit vehicles at each stop
T_A	automobile users generalized cost
T_C	maximum delay that auto users experience in the network
T_D	total delay that auto users experience in the network
T_E	total earliness that auto users experience in the network
T_e	length of earliness period
T_{el}	total penalty cost transit users consider for deviating from their wished schedules
T_F	total free-flow travel time that auto users experience in the network
T_I	total time that users spend in transit
T_l	length of lateness period
T_L	total lateness that auto users experience in the network
T_m	length of the study period
T_P	equivalent cost of fare that transit users pay to user the service
T_T	transit users generalized cost
T_Q	length of peak period
T_U	users generalized cost
T_W	total access and waiting time that transit users experience in the network
v	network average speed
v_F	arterial network free-flow speed
v_T	commercial speed of the transit vehicles
v_w	transit users walking speed
VOT	users' value of time
$W(t)$	cumulative distribution of the wished schedules of automobile users
α_E^A	automobile air pollution cost conversation factor for consuming a volume unit of fuel
α_E^T	transit vehicles air pollution cost conversation factor for consuming a volume unit of fuel
α_C	cost associated with constructing a unit of mile of the new infrastructure for the transit system
α_M	cost associated with one vehicle for the agency
α_{VMT}	cost of a vehicle mile travel in the network for the agency
α_{VHT}	cost of a vehicle hour travel in the network for the agency
β_C	relative importance factor of users cost in the total generalized cost function of the system
β_E	relative importance factor of emission cost in the total generalized cost function of the system
β_T	relative importance factor of operating cost in the total generalized cost function of the system
δ	average number of transfers between transit lines that users make in their trips
φ	length of the horizontal transit lines in the grid network
γ	choice model constant parameter
μ	network outflow
$\bar{\mu}$	equivalent capacity of the network in the bottleneck model
μ_m	network capacity
ρ	network allocation factor
τ_F	off-peak travel time
$\tau_D(N)$	delay that auto user N experiences in the network
$\tau_E(N)$	earliness that auto user N experiences in the network
$\tau_L(N)$	lateness that auto user N experiences in the network
$\tau_A(N)$	generalized cost that auto users N experiences in the network
τ_A^{avg}	average automobile users generalized cost
τ_T^{avg}	average transit users generalized cost
τ_D^{avg}	average delay that automobile users experience in the network
τ_S	travel time of the transit system between two consecutive stops
θ	transit vehicle to automobile conversion factor
θ_T	angular radial-line spacing
θ_N	angular radial-arterial spacing
Γ	departure time of the auto user who arrives to the network at time t
ϑ	penalty factor for the time it takes the transit users to transfer between the transit lines
ω	inverse Z-shaped curve constant slope

References

- Affum, J.K., Brown, A.L., Chan, Y.C., 2003. Integrating air pollution modelling with scenario testing in road transport planning: the TRAEMS approach. *Sci. Total Environ.* 312 (1), 1–14.
- Ahn, K., Rakha, H., Trani, A., Van Aerde, M., 2002. Estimating vehicle fuel consumption and emissions based on instantaneous speed and acceleration levels. *J. Transp. Eng.* 128 (2), 182–190.
- Alzuwayer, B., Abdelhamid, M., Pisu, P., Giovenco, P., Venhovens, P., 2014. Modeling and simulation of a series hybrid CNG vehicle. *SAE Int. J. Altern. Powertrains* 3 (2014-01-1802), 20–29.
- Amirgholy, M., Rezaeestakhruie, H., Poorzahedy, H., 2015. Multi-objective cordon price design to control long run adverse traffic effects in large urban areas. *NETNOMICS: Econ. Res. Electron. Network.* 16 (1–2), 1–52.

- Amirgholy, M., Gonzales, E.J., 2016. Demand responsive transit systems with time-dependent demand: user equilibrium, system optimum, and management strategy. *Transp. Res. Part B: Methodol.* 92, 234–252.
- Amirgholy, M., Gonzales, E.J., 2017. Analytical equilibrium of bicriterion choices with heterogeneous user preferences: application to the morning commute problem. *Transportmetrica B*, 1–33.
- Amirgholy, M., Gao, H.O., submitted for publication. Modeling dynamics of congestion in urban networks using macroscopic fundamental diagram: user equilibrium, system optimum, and pricing strategies. *Transp. Res. Part B: Methodol.* (submitted for publication).
- Arnott, R., de Palma, A., Lindsey, R., 1988. Schedule delay and departure time decisions with heterogeneous commuters. *Transp. Res. Rec.* 1197, 56–67.
- Arnott, R., de Palma, A., Lindsey, R., 1992. Route choice with heterogeneous drivers and group-specific congestion costs. *Region. Sci. Urban Econ.* 22 (1), 71–102.
- Arnott, R., de Palma, A., Lindsey, R., 1994. The welfare effects of congestion tolls with heterogeneous commuters. *J. Transp. Econ. Policy* 28 (2), 139–161.
- Arnott, R., 2013. A bathtub model of downtown traffic congestion. *J. Urban Econ.* 76, 110–121.
- Badia, H., Estrada, M., Robusté, F., 2014. Competitive transit network design in cities with radial street patterns. *Transp. Res. Part B: Methodol.* 59, 161–181.
- Bagherian, M., Mesbah, M., Ferreira, L., Charles, P., Khalilikhah, M., 2016. In: *Transportation Research Board 94th Annual Meeting*, No. 15-1146.
- Byrne, B.F., Vuchic, V.R., 1972. Public transportation line positions and headways for minimum cost. *Traffic Flow Transp.*
- Byrne, B.F., 1975. Public transportation line positions and headways for minimum user and system cost in a radial case. *Transp. Res.* 9 (2–3), 97–102.
- Chebbi, O., Chaouachi, J., 2016. Reducing the wasted transportation capacity of personal rapid transit systems: an integrated model and multi-objective optimization approach. *Transp. Res. Part E* 89, 236–258.
- Chen, H., Gu, W., Cassidy, M.J., Daganzo, C.F., 2015a. Optimal transit service atop ring-radial and grid street networks: a continuum approximation design method and comparisons. *Transp. Res. Part B: Methodol.* 81, 755–774.
- Chen, H., Nie, Y.M., Yin, Y., 2015b. Optimal multi-step toll design under general user heterogeneity. *Transp. Res. Procedia* 7, 341–361.
- Chien, S., Spasovic, L., Elefantiotis, S., Chhonkar, R., 2001. Evaluation of feeder bus systems with probabilistic time-varying demands and nonadditive time costs. *Transp. Res. Rec.: J. Transp. Res. Board* 1760, 47–55.
- Christofa, E., Skabardonis, A., 2011. Traffic signal optimization with application of transit signal priority to an isolated intersection. *Transp. Res. Rec.: J. Transp. Res. Board* 2259, 192–201.
- Christofa, E., Papamichail, I., Skabardonis, A., 2013. Person-based traffic responsive signal control optimization. *IEEE Trans. Intell. Transp. Syst.* 14 (3), 1278–1289.
- Daganzo, C.F., 1985. The uniqueness of a time-dependent equilibrium distribution of arrivals at a single bottleneck. *Transp. Sci.* 19 (1), 29–37.
- Daganzo, C.F., Smilowitz, K.R., 2004. Bounds and approximations for the transportation problem of linear programming and other scalable network problems. *Transp. Sci.* 38 (3), 343–356.
- Daganzo, C.F., 2005. *Logistics Systems Analysis*. Springer Science & Business Media.
- Daganzo, C.F., 2007. Urban gridlock: macroscopic modeling and mitigation approaches. *Transp. Res. Part B* 41 (1), 49–62.
- Daganzo, C.F., Geroliminis, N., 2008. An analytical approximation for the macroscopic fundamental diagram of urban traffic. *Transp. Res. Part B: Methodol.* 42 (9), 771–781.
- Daganzo, C.F., 2010a. Structure of competitive transit networks. *Transp. Res. Part B: Methodol.* 44 (4), 434–446.
- Daganzo, C.F., 2010b. Public transportation systems: basic principles of system design. *Oper. Plann. Real-Time Control*.
- Daganzo, C.F., 2012. On the design of public infrastructure systems with elastic demand. *Transp. Res. Part B: Methodol.* 46 (9), 1288–1293.
- Daganzo, C.F., Gayah, V.V., Gonzales, E.J., 2012. The potential of parsimonious models for understanding large scale transportation systems and answering big picture questions. *EURO J. Transp. Logist.* 1 (1–2), 47–65.
- Daganzo, C.F., 2013. System optimum and pricing for the day-long commute with distributed demand, autos and transit. *Transp. Res. Part B: Methodol.* 55, 98–117.
- Du, J., Rakha, H., Gayah, V.V., 2015a. Deriving macroscopic fundamental diagrams from probe data: issues and proposed solutions. *Transp. Res. Part C: Emerg. Technol.*
- Du, J., Rakha, H., Gayah, V., 2015b. Design and evaluation of network control strategies using the macroscopic fundamental diagram. In: *2015 IEEE 18th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, pp. 119–124.
- De Vlieger, I., 1997. On board emission and fuel consumption measurement campaign on petrol-driven passenger cars. *Atmos. Environ.* 31 (22), 3753–3761.
- Estrada, M., Roca-Riu, M., Badia, H., Robusté, F., Daganzo, C.F., 2011. Design and implementation of efficient transit networks: procedure, case study and validity test. *Transp. Res. Part A: Policy Pract.* 45 (9), 935–950.
- Fosgerau, M., Small, K.A., 2013. Hypercongestion in downtown metropolis. *J. Urban Econ.* 76, 122–134.
- Fosgerau, M., 2015. Congestion in the bathtub. *Econ. Transp.* 4 (4), 241–255.
- Frey, H.C., Roupail, N.M., Zhai, H., Farias, T.L., Gonçalves, G.A., 2007. Comparing real-world fuel consumption for diesel-and hydrogen-fueled transit buses and implication for emissions. *Transp. Res. Part D: Transp. Environ.* 12 (4), 281–291.
- Gayah, V.V., Daganzo, C.F., 2011. Clockwise hysteresis loops in the macroscopic fundamental diagram: an effect of network instability. *Transp. Res. Part B* 45 (4), 643–655.
- Gayah, V., Dixit, V., 2013. Using mobile probe data and the macroscopic fundamental diagram to estimate network densities: tests using microsimulation. *Transp. Res. Rec.: J. Transp. Res. Board* 2390, 76–86.
- Gayah, V.V., Gao, X.S., Nagle, A.S., 2014. On the impacts of locally adaptive signal control on urban network stability and the macroscopic fundamental diagram. *Transp. Res. Part B: Methodol.* 70, 255–268.
- Geroliminis, N., Daganzo, C.F., 2007. Macroscopic modeling of traffic in cities. In: *Transportation Research Board 86th Annual Meeting*, No. 07-0413.
- Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. *Transp. Res. Part B: Methodol.* 42 (9), 759–770.
- Geroliminis, N., Levinson, D.M., 2009. Cordon pricing consistent with the physics of overcrowding. In: *Transportation and Traffic Theory 2009: Golden Jubilee*. Springer, US, pp. 219–240.
- Geroliminis, N., Sun, J., 2011. Properties of a well-defined macroscopic fundamental diagram for urban traffic. *Transp. Res. Part B: Methodol.* 45 (3), 605–617.
- Geroliminis, N., Haddad, J., Ramezani, M., 2013. Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: a model predictive approach. *IEEE Trans. Intell. Transp. Syst.* 14 (1), 348–359.
- Geroliminis, N., Zheng, N., Ampountolas, K., 2014. A three-dimensional macroscopic fundamental diagram for mixed bi-modal urban networks. *Transp. Res. Part C: Emerg. Technol.* 42, 168–181.
- Gonzales, E.J., Geroliminis, N., Cassidy, M.J., Daganzo, C.F., 2010. On the allocation of city space to multiple transport modes. *Transp. Plann. Technol.* 33 (8), 643–656.
- Gonzales, E., Chavis, C., Li, Y., Daganzo, C.F., 2011. Multimodal transport in Nairobi, Kenya: insights and recommendations with a macroscopic evidence-based model. In: *Transportation Research Board 90th Annual Meeting*, No. 11-3045.
- Gonzales, E.J., Daganzo, C.F., 2012. Morning commute with competing modes and distributed demand: user equilibrium, system optimum, and pricing. *Transp. Res. Part B: Methodol.* 46 (10), 1519–1534.
- Gonzales, E.J., Daganzo, C.F., 2013. The evening commute with cars and transit: duality results and user equilibrium for the combined morning and evening peaks. *Transp. Res. Part B: Methodol.* 57, 286–299.
- Guler, S.I., Cassidy, M.J., 2012. Strategies for sharing bottleneck capacity among buses and cars. *Transp. Res. Part B: Methodol.* 46 (10), 1334–1345.
- Haddad, J., Geroliminis, N., 2012. On the stability of traffic perimeter control in two-region urban cities. *Transp. Res. Part B: Methodol.* 46 (9), 1159–1176.

- Haddad, J., Ramezani, M., Geroliminis, N., 2013. Cooperative traffic control of a mixed network with two urban regions and a freeway. *Transp. Res. Part B: Methodol.* 54, 17–36.
- Haddad, J., Shraiber, A., 2014. Robust perimeter control design for an urban region. *Transp. Res. Part B: Methodol.* 68, 315–332.
- Hendrickson, C., Kocur, G., 1981. Schedule delay and departure time decisions in a deterministic model. *Transp. Sci.* 15 (1), 62–77.
- Henderson, J.V., 1974. Road congestion: a reconsideration of pricing theory. *J. Urban Econ.* 1, 346–355.
- Henderson, J.V., 1977. *Economic Theory and the Cities*. Academic Press, New York (Chapter 8).
- Henderson, J.V., 1981. The economics of staggered work hours. *J. Urban Econ.* 9, 349–364.
- Holroyd, E.M., 1967. The optimum bus service: a theoretical model for a large uniform urban area. In: *Proceedings of the Third International Symposium on the Theory of Traffic Flow*.
- Huang, D., Liu, Z., Liu, P., Chen, J., 2016. Optimal transit fare and service frequency of a nonlinear origin-destination based fare structure. *Transp. Res. Part E* 96, 1–19.
- Ji, Y., Geroliminis, N., 2012. On the spatial partitioning of urban transportation networks. *Transp. Res. Part B: Methodol.* 46 (10), 1639–1656.
- Ji, Y., Luo, J., Geroliminis, N., 2014. Empirical observations of congestion propagation and dynamic partitioning with probe data for large-scale systems. *Transp. Res. Rec.: J. Transp. Res. Board* 2422, 1–11.
- Keyvan-Ekbatani, M., Kouvelas, A., Papamichail, I., Papageorgiou, M., 2012. Exploiting the fundamental diagram of urban networks for feedback-based gating. *Transp. Res. Part B: Methodol.* 46 (10), 1393–1403.
- Khalilikhah, M., Habibian, M., Heaslip, K., 2016. Acceptability of increasing petrol price as a TDM pricing policy: a case study in Tehran. *Transp. Policy* 45, 136–144.
- Lajunen, A., 2014. Fuel economy analysis of conventional and hybrid heavy vehicle combinations over real-world operating routes. *Transp. Res. Part D: Transp. Environ.* 31, 70–84.
- Laval, J.A., Castrillón, F., 2015. Stochastic approximations for the macroscopic fundamental diagram of urban networks. *Transp. Res. Procedia* 7, 615–630.
- Leclercq, L., Chiabaut, N., Trinquier, B., 2014. Macroscopic fundamental diagrams: a cross-comparison of estimation methods. *Transp. Res. Part B: Methodol.* 62, 1–12.
- Lindsey, R., 2004. Existence, uniqueness, and trip cost function properties of user equilibrium in the bottleneck model with multiple user classes. *Transp. Sci.* 38 (3), 293–314.
- Liu, Y., Nie, Y.M., 2011. Morning commute problem considering route choice, user heterogeneity and alternative system optima. *Transp. Res. Part B* 45 (4), 619–642.
- Liu, Y., Nie, Y., Hall, J., 2015. A semi-analytical approach for solving the bottleneck model with general user Heterogeneity. *Transp. Res. Part B* 71, 56–70.
- Mahmassani, H.S., Saberi, M., Zockaie, A., 2013. Urban network gridlock: theory, characteristics, and dynamics. *Transp. Res. Part C: Emerg. Technol.* 36, 480–497.
- Mazloumian, A., Geroliminis, N., Helbing, D., 2010. The spatial variability of vehicle densities as determinant of urban network capacity. *Philos. Trans. Roy. Soc. A: Math., Phys. Eng. Sci.* 368 (1928), 4627–4647.
- McKain, D.L., Clark, N., Balon, T.H., Moynihan, P.J., Lynch, S.A., Webb, T.C., 2000. Characterization of Emissions from Hybrid-electric and Conventional Transit Buses (No. 2000-01-2011). SAE Technical Paper.
- Nagle, A., Gayah, V., 2014. Accuracy of networkwide traffic states estimated from mobile probe data. *Transp. Res. Rec.: J. Transp. Res. Board* 2421, 1–11.
- Nesamani, K.S., Chu, L., McNally, M.G., Jayakrishnan, R., 2007. Estimation of vehicular emissions by capturing traffic variations. *Atmos. Environ.* 41 (14), 2996–3008.
- Newell, G.F., 1987. The morning commute for nonidentical travelers. *Transp. Sci.* 21 (2), 74–88.
- Newell, G.F., 1979. Some issues relating to the optimal design of bus routes. *Transp. Sci.* 13 (1), 20–35.
- Nourbakhsh, S.M., Ouyang, Y., 2012. A structured flexible transit system for low demand areas. *Transp. Res. Part B* 46 (1), 204–216.
- Nourinejad, M., Roorda, M.J., 2016. A continuous approximation model for the fleet composition problem on the rectangular grid. *OR Spectrum*, 1–29.
- Oh, Y., Park, J., Lee, J., Do Eom, M., Park, S., 2014. Modeling effects of vehicle specifications on fuel economy based on engine fuel consumption map and vehicle dynamics. *Transp. Res. Part D: Transp. Environ.* 32, 287–302.
- Olsson, L., 1994. Motor vehicle air pollution control in Sweden. *Sci. Total Environ.* 146, 27–34.
- Post, K., Kent, J.H., Tomlin, J., Carruthers, N., 1984. Fuel consumption and emission modelling by power demand and a comparison with other models. *Transp. Res. Part A: Gener. Rel.* 18 (3), 191–213.
- Qian, Z., Zhang, H.M., 2011. The morning commute problem with heterogeneous travelers: the case of continuously distributed parameters. *Transportmetrica* 9 (2), 1–26.
- Rakha, H., Ding, Y., 2003. Impact of stops on vehicle fuel consumption and emissions. *J. Transp. Eng.* 129 (1), 23–32.
- Ramezani, M., Haddad, J., Geroliminis, N., 2015. Dynamics of heterogeneity in urban networks: aggregated traffic modeling and hierarchical control. *Transp. Res. Part B: Methodol.* 74, 1–19.
- Saidi, S., Wirasinghe, S.C., Kattan, L., 2016. Long-term planning for ring-radial urban rail transit networks. *Transp. Res. Part B: Methodol.* 86, 128–146.
- Sbayti, H., El-Fadel, M., Kaysi, I., 2002. Effect of roadway network aggregation levels on modeling of traffic-induced emission inventories in Beirut. *Transp. Res. Part D: Transp. Environ.* 7 (3), 163–173.
- Sharma, N., Gangopadhyay, R., Dhyani, R., 2010. Methodology for estimation of CO2 reduction from mass rapid transit system (MRTS) project. *J. Sci. Ind. Res.* 69, 586–593.
- Sahinidis, N.V., 1996. BARON: a general purpose global optimization software package. *J. Global Optim.* 8 (2), 201–205.
- Shabanpour, R., Golshani, N., Derrible, S., Mohammadian, A.K., Miralinaghi, M., 2017a. A joint discrete-continuous model of travel mode and departure time choices. *Transp. Res. Rec.: J. Transp. Res. Board*. <http://dx.doi.org/10.3141/2669-05>.
- Shabanpour, R., Javanmardi, M., Fasihzaman, M., Miralinaghi, M., Mohammadian, A., 2017b. Investigating the applicability of ADAPTS activity-based model in air quality analysis. *Travel Behav. Soc.* <http://dx.doi.org/10.1016/j.tbs.2017.02.004>.
- Silva, H., Lindsey, R., De Palma, A., Van den Berg, V.A., 2014. On the Existence and Uniqueness of Equilibrium in the Bottleneck Model with Atomic Users.
- Silva, C., Ross, M., Farias, T., 2009. Evaluation of energy consumption, emissions and cost of plug-in hybrid vehicles. *Energy Convers. Manage.* 50 (7), 1635–1643.
- Sivakumaran, K., Li, Y., Cassidy, M., Madanat, S., 2014. Access and the choice of transit technology. *Transp. Res. Part A: Policy Pract.* 59, 204–221.
- Small, K.A., Chu, X., 2003. Hypercongestion. *J. Transp. Econ. Policy*, 319–352.
- Smith, M.J., 1984. The existence of a time-dependent equilibrium distribution of arrivals at a single bottleneck. *Transp. Sci.* 18 (4), 385–394.
- Tawarmalani, M., Sahinidis, N.V., 2004. Global optimization of mixed-integer nonlinear programs: a theoretical and computational study. *Math. Program.* 99 (3), 563–591.
- Tian, L.J., Yang, H., Huang, H.J., 2013. Tradable credit schemes for managing bottleneck congestion and modal split with heterogeneous users. *Transp. Res. Part E: Logist. Transp. Rev.* 54, 1–13.
- Sobrinho, N., Monzon, A., Hernandez, S., 2016. Reduced carbon and energy footprint in highway operations: the Highway Energy Assessment (HERA) methodology. *Networks Spat. Econ.* 16 (1), 395–414.
- Tong, H.Y., Hung, W.T., Cheung, C.S., 2000. On-road motor vehicle emissions and fuel consumption in urban driving conditions. *J. Air Waste Manage. Assoc.* 50 (4), 543–554.
- Van den Berg, V., Verhoef, E.T., 2011a. Congestion tolling in the bottleneck model with heterogeneous values of time. *Transp. Res. Part B* 45 (1), 60–70.
- Van den Berg, V., Verhoef, E.T., 2011b. Winning or losing from dynamic bottleneck congestion pricing?: The distributional effects of road pricing with heterogeneity in values of time and schedule delay. *J. Public Econ.* 95 (7), 983–992.
- Vaughan, R., 1986. Optimum polar networks for an urban bus system with a many-to-many travel demand. *Transp. Res. Part B: Methodol.* 20 (3), 215–224.

- Vickrey, W., 1969. Congestion theory and transport investment. *Am. Econ. Rev.* 56, 251–260.
- Wang, H., Fu, L., Lin, X., Zhou, Y., Chen, J., 2009. A bottom-up methodology to estimate vehicle emissions for the Beijing urban area. *Sci. Total Environ.* 407 (6), 1947–1953.
- Wirasinghe, S.C., Hurdle, V.F., Newell, G.F., 1977. Optimal parameters for a coordinated rail and bus transit system. *Transp. Sci.* 11 (4), 359–374.
- Woensel, T.O.M., Creten, R., Vandaele, N., 2001. Managing the environmental externalities of traffic logistics: the issue of emissions. *Prod. Oper. Manage.* 10 (2), 207–223.
- Wu, W.X., Huang, H.J., 2015. An ordinary differential equation formulation of the bottleneck model with user heterogeneity. *Transp. Res. Part B* 81, 34–58.
- Xiao, L.L., Huang, H.J., Liu, R., 2015. Tradable credit scheme for rush hour travel choice with heterogeneous commuters. *Adv. Mech. Eng.* 7 (10), 1687814015612430.
- Xiao, F., Qian, Z., Zhang, H.M., 2011. The morning commute problem with coarse toll and nonidentical commuters. *Network Spat. Econ.* 11 (2), 343–369.
- Xie, D.F., Wang, D.Z., Gao, Z.Y., 2016. Macroscopic analysis of the fundamental diagram with inhomogeneous network and instable traffic. *Transportmetr. A: Transp. Sci.* 12 (1), 20–42.
- Yu, Y., Machemehl, R.B., Xie, C., 2015. Demand-responsive transit circulator service network design. *Transp. Res. Part E* 76, 160–175.