Transportation (2020) 47:2757–2776
https://doi.org/10.1007/s11116-019-09990-w

Can passenger flow distribution be estimated solely based on network properties in public transport systems?

Ding Luo[1] · **Oded Cats**[1] · **Hans van Lint**[1]

Published online: 6 March 2019

Abstract
We present a pioneering investigation into the relation between passenger flow distribution and network properties in public transport systems. The methodology is designed in a reverse engineering fashion by utilizing passively measured passenger flow dynamics over the entire network. We quantify the properties of public transport networks using a range of centrality indicators in the topological representations of public transport networks with both infrastructure and service layers considered. All the employed indicators, which originate from complex network science, are interpreted in the context of public transport systems. Regression models are further developed to capture the correlative relation between passenger flow distribution and several centrality indicators that are selected based on the correlation analysis. The primary finding from the case study on the tram networks of The Hague and Amsterdam is that the selected network properties can indeed be used to approximate passenger flow distribution in public transport systems to a reasonable extent. Notwithstanding, no causality is implied, as the correlation may also reflect how well the supply allocation caters for the underlying demand distribution. The significance and relevance of this study stems from two aspects: (1) the unraveled relation provides a parsimonious alternative to existing passenger assignment models that require many assumptions on the basis of limited data; (2) the resulting model offers efficient quick-scan decision support capabilities that can help transport planners in tactical planning decisions.

Keywords Public transport systems · Passenger flow distribution · Network properties · Topology · Centrality · Complex network science

Introduction

Estimation and prediction of passenger flow distribution is one of the most significant topics in the field of public transport (PT) research given its critical role in assisting planning and management. The conventional approach, like that in the road traffic research, is to develop passenger assignment models which take demand profiles—typically in the

✉ Ding Luo
d.luo@tudelft.nl

[1] Department of Transport and Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology, P.O. Box 5048, 2600 GA Delft, The Netherlands

form of origin–destination matrices—as input and then distribute the demand across the network (Ortúzar and Willumsen 2011). These models are normally referred to as *transit assignment models* in the transport research community, and their core pertains to modeling travelers' route choices in PT systems as functions of network conditions and travel preferences (Liu et al. 2010). Two types of static equilibrium transit assignment models have been mostly developed over the past decades, namely the frequency-based and schedule-based. The major distinction between them lies in the representation of public transport networks (PTNs) given their substantial impact on the passenger loading procedure (Gentile et al. 2016). More specifically, the frequency-based approach represents PTNs at the route-level with corresponding frequencies (e.g., Nguyen and Pallottino 1988; Spiess and Florian 1989; Cepeda et al. 2006; Schmöcker et al. 2011), while the schedule-based one enables a more detailed representation of time-dependent specific vehicle runs (e.g., Nuzzolo et al. 2001; Zhang et al. 2010).

Notwithstanding the continuous development of transit assignment models, other possibilities for understanding and further modeling the passenger flow distribution in PT systems to a network-wide extent have remained underexplored. Presumably, this is a result of the longstanding data scarcity in the field. Under this "data-poor and assumption-rich" situation (Vlahogianni et al. 2015), the conventional modeling approach has undoubtedly provided the most feasible solution to this challenging problem. Nonetheless, as the capability in measuring the PT passenger flow dynamics in a large spatiotemporal scale becomes increasingly available owing to emerging PT demand data sources, such as the automatic fare collection data (Pelletier et al. 2011), it is now worth investigating whether there can be alternative ways to model the passenger flow distribution in PT systems.

This study hence examines a research question: Can passenger flow distribution be estimated solely based on network properties in PT systems? While the answer to this question looks apparent, it has not been empirically investigated to a sufficient extent. One can make an underlying assumption that transport network properties should of course correspond to passenger flow distribution since networks are supposedly designed to efficiently accommodate prevailing demand patterns in PT systems (van Nes et al. 1988). However, it shall be stressed that a range of other factors, such as travelers' behavior, historical network development and physical constraints, also have non-negligible influences on demand and network structure in any transport systems. In fact, the discussion about whether traffic flows can be approximated by network properties in urban street networks has lasted for decades among urban planning researchers (e.g., Hillier et al. 1993; Penn et al. 1998; Turner 2007; Jiang and Liu 2009; Kazerani 2009; Gao et al. 2013). Recent evidence was provided by Gao et al. (2013) based on the traffic volume derived from the GPS-enabled taxi trajectory data from a Chinese city. Their study concludes that the betweenness centrality, which has been commonly employed as a local indicator of network properties, is not a good predictive variable for urban traffic flow and the gap can be explained by the spatial heterogeneity of human activities and the distance-decay law. In addition, a limited amount of research attempts have also been made by scholars from various fields in the past few years to examine the relation between network properties—mostly limited to the betweenness centrality—and the traffic flows in urban road traffic systems. (e.g., Altshuler et al. 2011; Puzis et al. 2013; Ye et al. 2016; Zhao et al. 2017; Wen et al. 2017; Akbarzadeh et al. 2017). No such comparable effort, however, has been made in the context of PT systems, which therefore necessitates dedicated investigations into the proposed research question above.

To this end, we conduct this study with the methodology developed in a reverse engineering fashion, which unravels the correlative relation between passenger flow distribution

and network properties in PT systems. Differing from previous studies, we examine a variety of network properties by considering the centrality indicators in different topological representations of PTNs. We show how concepts originating from the domain of complex network science can be applied and interpreted in the context of PT systems. We further apply the proposed methodology to two real-world tram networks in The Netherlands, i.e., The Hague and Amsterdam, where passenger flow observations are available. Regression models capturing the correlation between passenger flow distribution and several centrality indicators are first developed using the data from The Hague, and are then evaluated for both networks separately. Note that no causality is implied by the models. It is the correlation—rather than the causation—between PTN properties and passenger flow distribution that is essentially investigated. Moreover, the unraveled relation and developed models have the potential to serve as a complementary tool for PT operations management, while it is inappropriate to apply them to the long-term passenger flow forecasting.

The remainder of this paper is organized as follows: second section displays the proposed methodology. Third section describes the case study networks and experimental setup, which is followed by the presentation of the results and discussion in fourth section. The conclusions are drawn in final section with some remarks on the future research directions.

Methodology

Overview

An overview of the research structure is shown in Fig. 1 with the workflow and components of the methodology sketched. In the beginning, the representation of PTNs is described in "Representation of public transport networks" section to lay the foundation. Then, "Independent variables: time-dependent centrality indicators of PTNs" and "Dependent variable: time-dependent passenger flow distribution" sections are respectively dedicated to illustrating the independent and dependent variables in this study, namely centrality indicators of PTNs and passenger flow distribution, both of which are considered in a time-dependent manner. The model development is later described in "Model development" section, followed by the model evaluation in "Model evaluation" section.

Representation of public transport networks

We first clarify that the term "public transport network" in this study is referred to as the combination of two layers: the infrastructure network (i.e., road and rail) and the service network superimposed on the physical layer (i.e., routes). We then define a PTN as a *directed graph* with a triple $G = (V, E, R)$, where V, E, R represent the set of *nodes, links, routes*, respectively. Each node $v \in V$ represents a stop, while each link $e \in E$ is defined by an ordered pair of nodes (u, v), where u and v, respectively, denote the *source* and *target* nodes. Each route $r \in R$ is characterized by an ordered sequence of stops $r = (v_{r,1}, v_{r,2}, \dots, v_{r,|r|})$ as well as an ordered sequence of links $r = (e_{r,1}, e_{r,2}, \dots, e_{r,|r|})$. Note that a link can be utilized by multiple routes, and the direction of a route is also distinguished. In addition, the stop in this definition refers to a service location which can contain more than one individual boarding and alighting spot in the operational network.

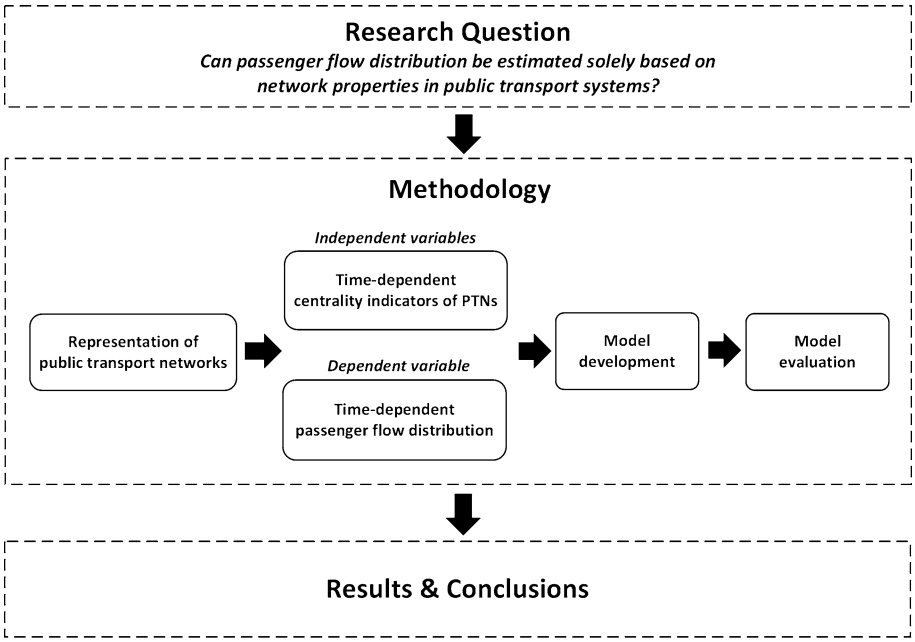

Fig. 1 Illustration of the overall research design and methodology

Based on the fundamental representation of PTNs, we further apply two topological representations, the **L**- and **P**-space (von Ferber et al. 2009), to characterize the topology of PTNs' two different layers, i.e., infrastructure and service. These topological networks, which can be represented by adjacency matrices, are suitable inputs for further analyses. As Fig. 2 illustrates, the **L**-space is a straightforward representation of PTNs' physical infrastructure. Each node represents a stop, and a link between two stops is formed if two stops are adjacent on an infrastructure segment (i.e. road or rail). Moreover, duplicate connections between nodes are not allowed. The **P**-space is constructed solely based on the service layer designed by PT operators/agencies, i.e., routes. The nodes in this space also represent stops, and two nodes are linked if they are served by at least one common route. In this sense the neighbors of a node in this space are all stops that can be reached without performing a transfer. In order to make the use of these two topological representations more informative in the context of this study, we replace the terms "**L**-space" and "**P**-space" with "*space-of-infrastructure*" and "*space-of-service*" in the remaining of this paper.

Further enrichment of the topological networks of PTNs is performed by adding link weights related to PT service attributes. The space-of-infrastructure is enriched in two ways, including the in-vehicle travel time as a type of link cost and vehicle frequency per time unit as a type of link importance. With common routes considered, the weight of a link's ultimate frequency is determined by summing up the frequencies of all the routes traversing it, i.e., labeling the link with the respective joint frequency, which is consistent with the definition of space-of-infrastructure representation. For the space-of-service, the expected waiting time for a PT vehicle during a given time slice is considered as a type of link cost, which is defined as half of the planned headway with joint vehicle frequency

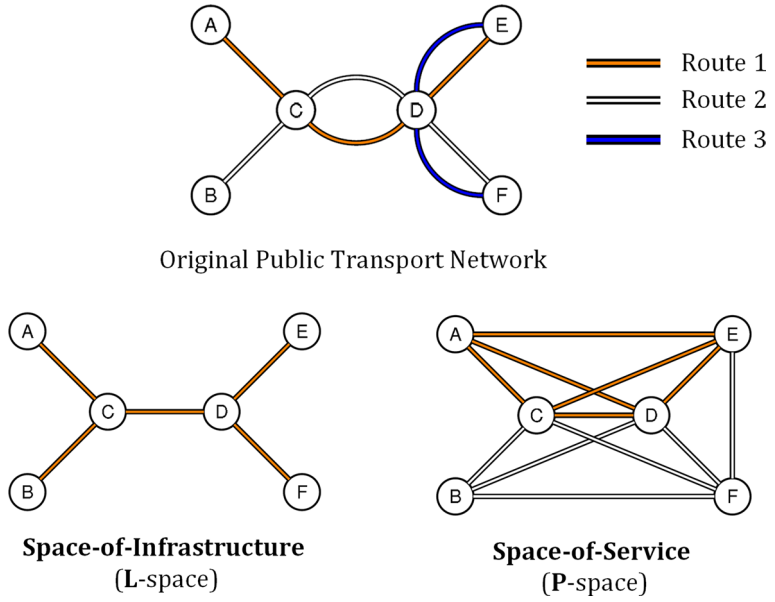

Fig. 2 Illustration of the **L**-space and **P**-space representations of the exemplary PTN on the top, which consists of three routes and six stops (adapted from von Ferber et al. 2009). The **L**-space essentially represents the infrastructure layout, while the **P**-space characterizes the PT service layer: stops that are directly linked require no transfer to reach each other. In order to make the use of these two topological representations more intuitive in the context of this study, we replace the term "**L**-space" and "**P**-space" with "*space-of-infrastructure*" and "*space-of-service*" in the remaining of this paper

between stop pairs considered. This definition is based on the assumption that (1) passenger arrival at stop is random in the context of urban high-frequency services, and (2) arrival times of vehicles serving different lines is independent, i.e. no systematic synchronization is performed in the context of urban high-frequency services. Both unweighted and weighted topological networks will be used in the following subsection.

Independent variables: time-dependent centrality indicators of PTNs

Since the introduction of the "centrality" concept by Bavelas (1948), a variety of network centrality indicators have been proposed in the past decades. In principal, all these indicators are designed to capture distinct aspects of what it means to be "central" in a network for individual nodes. Based on this concept, this study employs several different centrality indicators for both space-of-infrastructure and space-of-service networks as the proxies of different properties of PTNs. The combination of different topological representations and centrality indicators enables a concise way to quantify a range of fundamental properties of PTNs. Moreover, some centrality indicators are computed in time-dependent weighted networks, which correspondingly reflect time-dependent characteristics of PTNs.

A summary of all the employed centrality indicators is shown in Table 1 and detailed descriptions are presented in the following subsections. General definitions of the selected centrality are first given, followed by their interpretation in different topological

Table 1 Summary of the centrality indicators used in this study

PTN representation	Notations	Centrality indicators	Weight	Weight attributes
Space-of-infrastructure	$\mathbf{d}^{\mathbf{L},+/-}$	In/out-degree	✗	–
	$\tilde{\mathbf{d}}^{\mathbf{L},+/-}$	In/out-degree	✓	Vehicle frequency
	$\mathbf{b}^{\mathbf{L}}$	Betweenness	✗	–
	$\tilde{\mathbf{b}}^{\mathbf{L}}$	Betweenness	✓	In-vehicle travel time
	$\mathbf{c}^{\mathbf{L},+/-}$	In/out-closeness	✗	–
	$\tilde{\mathbf{c}}^{\mathbf{L},+/-}$	In/out-closeness	✓	In-vehicle travel time
Space-of-service	$\mathbf{d}^{\mathbf{P},+/-}$	In/out-degree	✗	–
	$\mathbf{b}^{\mathbf{P}}$	Betweenness	✗	–
	$\tilde{\mathbf{b}}^{\mathbf{P}}$	Betweenness	✓	Waiting time

representations of PTNs. In addition, all the centrality indicators of the nodes are scaled by having the division over the sum for comparability and transferability reasons.

In/out-degree centrality

For unweighted directed networks, the *in/out-degree* centrality is an indicator that determines the importance of a node based on the number of links connected to it in an inbound/outbound manner. This indicator can be further extended by adding weights to the network as proposed by Barrat et al. (2003), which is coined by them as *strength*. We stick to the term "degree in weighted networks" in the remaining of this study for the consistency with other indicators. The calculation of the in/out degree centrality in a weighted network can be based on the adjacency matrix A of it shown as follows:

$$\tilde{d}_i^+ = \sum_j w_{ji} A_{ji} \tag{1}$$

$$\tilde{d}_i^- = \sum_j w_{ij} A_{ij} \tag{2}$$

where $\tilde{d}_i^+$ and $\tilde{d}_i^-$ respectively denotes the in- and out-degree centrality of node i in a weighted network. w_{ij} denotes the value of weight of the corresponding link. When there is no weight considered , namely $w_{ij} = 1$, the indicators are degraded to the in- and out-degree centrality of node i in an unweighted network, denoted by d_i^+ and d_i^-.

- $\mathbf{d}^{\mathbf{L},+/-}$: In/out-degree centrality in the **unweighted** space-of-infrastructure network
  This indicator corresponds to the number of road or rail links that directly lead in or out of a given stop. It thus directly relates to the underlying physical infrastructure of PTNs.
- $\tilde{\mathbf{d}}^{\mathbf{L},+/-}$: In/out-degree centrality in the **weighted** space-of-infrastructure network
  Links are weighted by the time-dependent vehicle frequency between two adjacent stops with all the routes considered. Hence, this indicator quantifies the scheduled service intensity in terms of PT vehicle flows.
- $\mathbf{d}^{\mathbf{P},+/-}$: In/out-degree centrality in the **unweighted** space-of-service network

This indicator measures the number of stops that can be reached without transfer for a given stop. It thus directly relates to the underlying service design of PTNs.

Betweenness centrality

The *betweenness* centrality is a widely used indicator that was initially proposed by Freeman (1977) for social network studies. It quantifies the importance of a node in a network by measuring the proportion of the shortest paths between all node pairs in the network that pass through it. Assuming that flow travels through a network along the shortest path, nodes that lie on many shortest paths will undertake a high proportion of traffic, thus becoming more central in the network. In this sense, such a node might play a significant role in the passage of traffic through the network. The definition of the betweenness centrality is given as follows:

$$b_i = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{3}$$

where b_i denotes the betweenness centrality of node i. σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(i)$ is the number of those paths that pass through i.

Computing the betweenness centrality involves searching for all the shortest paths between node pairs. In this study, instead of leveraging on one single betweenness centrality indicator, the betweenness centrality indicators in both topological representations i.e. the space-of-infrastructure and the space-of-service—are considered. The major advantage is that through their inclusion, we are able to directly estimate the contribution of each cost component: in-vehicle time (weighted space-of-infrastructure), number of transfers (unweighted space-of-service) and waiting time (weighted space-of-service). Consequently, their contributions to model prediction power are disentangled without pre-specifying any behavioral trade-offs in this process. An additional advantage is that the computational burden of the betweenness centrality is also greatly relieved in this way since the algorithm proposed by Brandes (2001) can be easily applied in our case. The betweenness centrality indicators in different topological networks are explained below.

- $\mathbf{b^L}$: Betweenness centrality in the **unweighted** space-of-infrastructure network
  The share of shortest paths that traverse a certain stop when path length is measured in terms of the number of stops traversed. Given some evidence (Guo 2011), this indicator may coincide with how travelers choose their routes in complex PTNs using the map provided by agencies/operators as a mean to approximate travel time.
- $\tilde{\mathbf{b}}^{\mathbf{L}}$: Betweenness centrality in the **weighted** space-of-infrastructure network
  With the network weighted by the in-vehicle travel time, this indicator corresponds to the share of shortest paths in terms of on-board travel time that traverse the respective stop. Note that no regard is made to line configuration and thus the number of transfers induced.
- $\mathbf{b^P}$: Betweenness centrality in the **unweighted** space-of-service network
  This indicator relates to the interchange (hub) function of the respective stop. It therefore pertains to one of the most important and unique properties of PT systems, namely transfers.
- $\tilde{\mathbf{b}}^{\mathbf{P}}$: Betweenness centrality in the **weighted** space-of-service network

The share of shortest paths measured in terms of the average waiting time that traverse a given stop. The path cost consists of the waiting time at the first stop for the route chosen and the waiting time at all subsequent transfer locations.

In/out-closeness centrality

The intuition of the *closeness* centrality is that two nodes in a network are maximally close—in a topological sense—if they share a direct connection, whereas two nodes that are only tied indirectly through many intermediate nodes are topologically distant (Bavelas 1950). Given this logic, the topological distance between two nodes can be defined as the number of links on the shortest path between them. Hence, a node becomes topologically central if it is able to interact with many network elements via only a few links, namely having a short average path length. More formally, the closeness centrality of a node can be defined as the inverse of its average shortest path length (Beauchamp 1965):

$$c_i = \frac{N-1}{\sum_{j \neq i} l_{ij}} \tag{4}$$

where c_i denotes the closeness centrality of node i. l_{ij} is the shortest path length, or topological distance, between nodes i and j. N is the number of nodes in the network. In directed networks, if we define that l_{ij} is the shortest path from node j to node i, then Eq. 4 depicts the closeness centrality according to the shortest paths that are incoming to node i, which is defined as the *in-closeness* centrality. Similarly, the *out-closeness* centrality is based on the paths outgoing from node i, in which case we would instead sum over l_{ji} for $j = 1, ..., N$ in Eq. 4. In weighted networks, the closeness centrality can be estimated by searching the shortest weighted path length between regions, where the weight of the path is determined by the sum of the link weights on that path.

- $\mathbf{c}^{\mathbf{L},+/-}$: In/out-closeness centrality in the **unweighted** space-of-infrastructure network
  This indicator quantifies the phenomenon that passengers originating from the topologically central stops can reach the others in the network with fewer intermediate ones.
- $\tilde{\mathbf{c}}^{\mathbf{L},+/-}$: In/out-closeness centrality in the **weighted** space-of-infrastructure network
  The weight is determined by the scheduled in-vehicle travel time, thus making the shortest path more related to the PT service.

The closeness centrality in the space-of-service network is not included in model development because it reflects a concept very similar to the one obtained through the degree centrality in the same space ($\mathbf{d}^{\mathbf{P},+/-}$), namely identifying the stops that are most reachable with the least number of transfers.

Dependent variable: time-dependent passenger flow distribution

The time-dependent passenger flow distribution at PT stops is leveraged as the dependent variable, denoted by $\mathbf{q}$. Here we define the passenger flow at a stop in PTNs as the sum of inflow, outflow and throughflow at this stop during specified time slices. Specifically, inflow and outflow respectively represent the amount of passengers entering (boarding)/exiting (alighting) the PT system at a stop, while throughflow represents the amount of passengers that pass through a stop without leaving PT vehicles. This definition of the passenger flow sufficiently characterizes how intensively the stops are used across the network. In addition, the absolute

passenger flows are converted into relative terms, i.e. flow share, at each stop divided by the sum of all stop flows across the network during the respective time slices. we do not attempt to directly predict absolute flow values based on scaled centrality indicators because the same centrality value may correspond to different contexts for different networks and time periods. Instead, we examine whether the distribution of passenger flows is correlated with service properties by considering each stop and time-period as a single observation. Absolute flow values are resorted by multiplying flow shares by the total passenger flow in the network.

Model development

The model development is performed in two steps, with the first being an exploratory analysis among variables based on the Pearson correlation coefficient, and the second being building regression models. The objective of the first step is to find out (1) which independent variables (centrality indicators) have higher correlation with the dependent variable (passenger flow distribution), and thus can be incorporated into the models to be developed; (2) the collinearity among independent variables. This is to ensure that variables that are mutually linearly correlated are not included in the models at the same time so that the developed models are as parsimonious as possible.

Following the exploratory analysis, we estimate regression models to capture the correlative relation between passenger flow distribution and network properties. Each observation in the dataset corresponds to the flow share associated with a given stop for a given time instance. Hence, the dataset contains (balanced) panel data, which are time-dependent and cross-sectional. Panel data regression models are thus applied. Let us denote the number of time periods for which each element (i.e. stop) i is observed as T_i. Panel data models are most useful when the outcome variable is expected to depend on explanatory variables which are not directly observable but correlated with the observed explanatory variables. If such omitted variables are time-invariable, panel data estimators allow to consistently estimate the effect of the observed explanatory variables. A general formulation of the panel data regression model with specific individual effects is presented below:

$$y_{it} = \alpha + \beta X_{it} + \mu_i + v_{it}, \quad i = 1, \dots, n, \quad t = 1, \dots, T_i \tag{5}$$

where μ_i represents the ith invariant time individual effect and $v_{it} \sim i.i.d(0, \theta_v^2)$ the disturbance. There are several different estimators (e.g., fixed effects, random effects, mixed effects, etc.) for panel data models based on different assumptions, of which more details can be found in relevant literature (e.g., Hsiao 2007). In this study, the *random effects (RE)* model is applied in order to relieve the loss of degree of freedom during the estimation, as the number of units in our case is quite large (hundreds of PT stops). In RE models, the individual-specific effect is assumed to be a random variable that is uncorrelated with the explanatory variables, i.e., $\text{Cov}(X_{it}, \mu_i) = 0$ and $\text{Cov}(X_{it}, v_{it}) = 0$ for all i and t. The model can be then formulated as:

$$y_{it} = \alpha + \beta X_{it} + \mu_{it}, \quad i = 1, \dots, n, \quad t = 1, \dots, T_i \tag{6}$$

where $\mu_{it} = \mu_i + v_{it}$ represents the error term that includes the ith invariant time individual effects μ_i and the disturbance v_{it}.

Model evaluation

Absolute passenger flows are used in the evaluation of the estimated models. These values, which are also time-dependent, are derived by multiplying the relative one (flow shares) that are obtained from the models by the total amount of flows in the network. Four evaluation measures, including the mean absolute error (MAE), weighted mean absolute error (WMAE), weighted absolute percentage error (MAPE) and weighted mean absolute percentage error (WMAPE). The motivation for taking into account the weighted measures—of which weights are determined by the magnitude of passenger flows at the corresponding stops—is that we want to reduce the bias caused by extreme error values at stops with low passenger flows. The applied measures are specified as below:

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \tag{7}$$

$$WMAE = \frac{\sum_{i=1}^{n} w_i \times |\hat{y}_i - y_i|}{\sum_{i=1}^{n} w_i} \tag{8}$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \tag{9}$$

$$WMAPE = \frac{100\%}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \times \left| \frac{\hat{y}_i - y_i}{y_i} \right| \tag{10}$$

where $\hat{y}_i$ and y_i denote the predicted and actual passenger flows of stop i, respectively. w_i represents the weight of stop i, namely the flow share. n denotes the total number of observations for the evaluation data set.

Studied networks and experimental setup

Networks and data

The tram networks of two Dutch cities—The Hague and Amsterdam—were used for this investigation given the rich data availability of the Dutch PT systems (van Oort et al. 2015). The data of the entire month of March, 2015 for The Hague, and that of the entire day of November 14th, 2017 for Amsterdam were leveraged. As Fig. 3 shows, automatic fare collection (AFC), automatic vehicle location (AVL), and general transit feed specification (GTFS) data were used as major sources to generate networks as well as highly aggregated spatiotemporal data sets of dependent and independent variables. The passenger flow distribution (dependent variable) were obtained from the PT vehicle trajectories with passenger loads (Luo et al. 2018). A summary of the basic properties of the two networks, including the number of nodes, (directional) routes, links in space-of-infrastructure and space-of-service networks, is presented in Table 2.

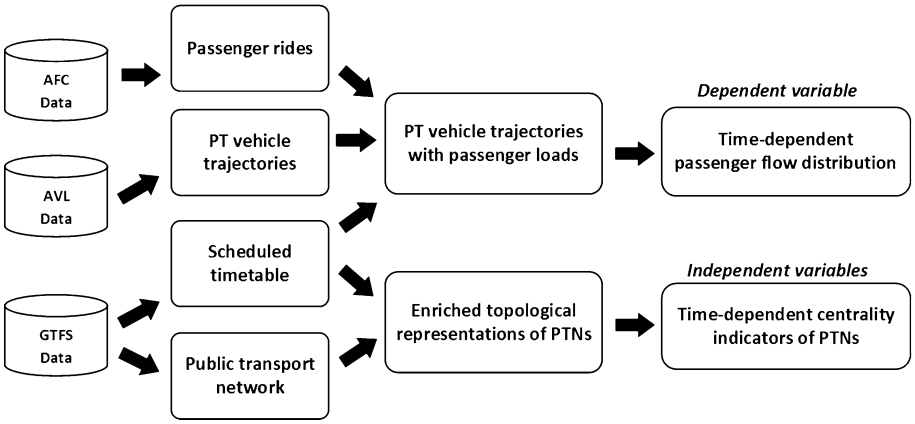

Fig. 3 Workflow of the data preparation

Table 2 Summary of the studied tram networks

Basic properties	The Hague	Amsterdam
Nodes	232	192
Directional routes	28	24
Links in space-of-infrastructure	520	418
Links in space-of-service	8901	6122

Experimental setup

For the experimental setup, we selected 20 working days with normal demand patterns (out of 1 month) for The Hague. 15 working days were further randomly selected for the model development, with the data aggregated on an hourly basis from 6 a.m. to 12 a.m. (18 time slices). The rest 5-day data set of The Hague and the 1-day data set of Amsterdam were utilized for the model evaluation.

Results and discussion

The results of the exploratory analysis on the two employed networks are first shown in the first subsection. The second subsection then presents the results of model estimation, followed by the model evaluation in the final subsection.

Exploratory analysis

To gain more intuition about the spatial distribution of passenger flow and centrality indicators in the studied networks, the visualizations of them for the weekday morning peak (7 a.m.–8 a.m.) are performed and presented in Figs. 4 (The Hague) and 5 (Amsterdam). Both size and color are used to make the distinction in magnitude remarkable. Out-degree and

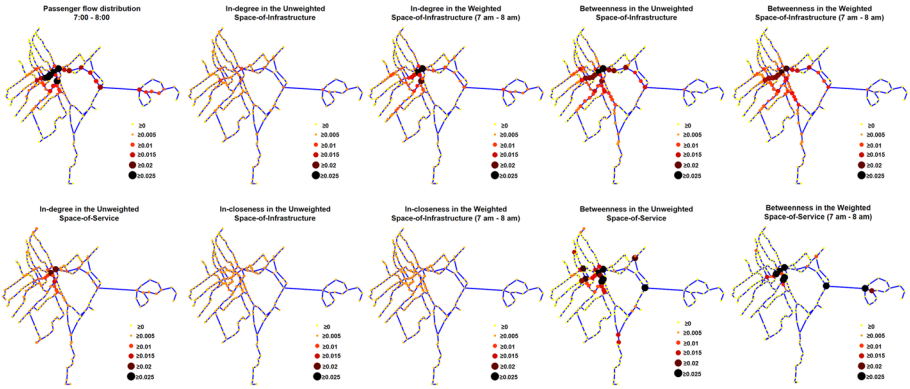

Fig. 4 Visualization of the passenger flow distribution and the employed centrality indicators for the weekday morning peak (7 a.m.–8 a.m.) of the tram network of The Hague

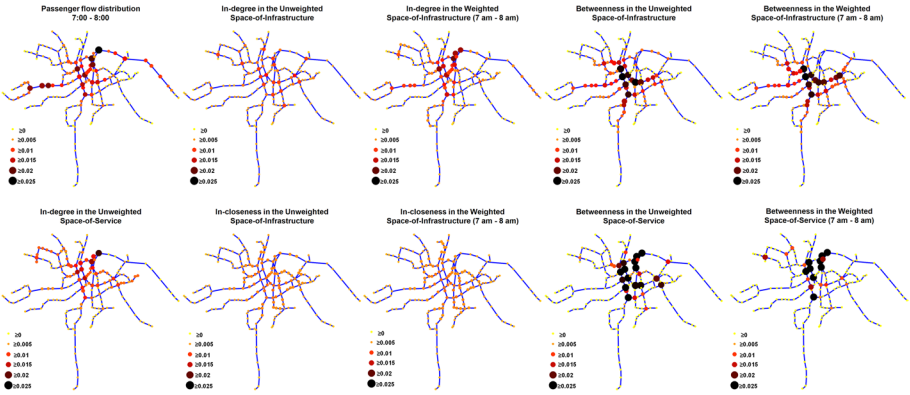

Fig. 5 Visualization of the passenger flow distribution and the employed centrality indicators for the weekday morning peak (7 a.m.–8 a.m.) of the tram network of Amsterdam

out-closeness are omitted as they display the same pattern as their counterparts. Through the visualizations, it can be seen that considerable amount of passenger flows are loaded in the central area of both networks, though it is also observable that some corridors used by commuters also undertake a significant amount of flows, such as the one from center to the east in The Hague, and two horizontal corridors in Amsterdam with one on the middle of west and the other on the top of east. We can further notice that the in-degree centrality in the weighted space-of-infrastructure ($\tilde{\mathbf{d}}^{\mathbf{L},+}$) and the betweenness centrality in both unweighted ($\mathbf{b}^{\mathbf{L}}$) and weighted ($\tilde{\mathbf{b}}^{\mathbf{L}}$) space-of-infrastructure mostly match the flow distribution pattern with clear distinctions among nodes across the networks. Some indicators, including the in-degree in the unweighted space-of-infrastructure ($\mathbf{d}^{\mathbf{L},+}$) and the in-closeness in both unweighted ($\mathbf{c}^{\mathbf{L},+}$) and weighted ($\tilde{\mathbf{c}}^{\mathbf{L},+}$) space-of-infrastructure, show rather plain patterns. Besides, the betweenness in the space-of-service ($\mathbf{b}^{\mathbf{P}}$ and $\tilde{\mathbf{b}}^{\mathbf{P}}$) makes the transfer locations in the networks really stand out.

The temporal variance in the dependent and independent variables for both networks is further displayed through the distribution plots in Fig. 6. Four different time slices are selected for each variable to demonstrate the distinction between peak (07:00–08:00 and

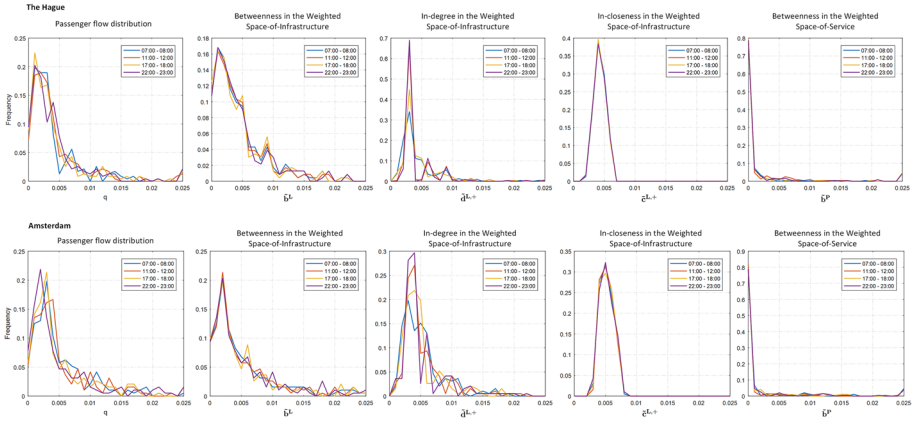

Fig. 6 Illustration of the temporal variation of the distributions of dependent and independent variables for both The Hague and Amsterdam tram networks. Four different time slices are selected to display mainly the difference between peak (07:00–08:00 and 17:00–18:00) and non-peak periods (11:00–12:00 and 22:00–23:00)

17:00–18:00) and non-peak (11:00–12:00 and 22:00–23:00) periods. For the dependent variable, i.e. share of passengers flow, a few nodes are traversed by a large proportion of the flows while the rest of them only share a small proportion. This pattern is persistent over all time periods and is observed for both networks. As for the independent variables, significant differences across the four time periods can be observed for $\tilde{\mathbf{d}}^{\mathbf{L},+}$ since it largely depends on the planned service frequency which varies over the day. Differences also hold for $\tilde{\mathbf{b}}^{\mathbf{L}}$, albeit to a lesser extent, due to different traffic conditions.

We further visualize the correlation coefficient matrices among all the variables for both networks as shown in Fig. 7. With the passenger flow distribution $\mathbf{q}$ placed at the first place

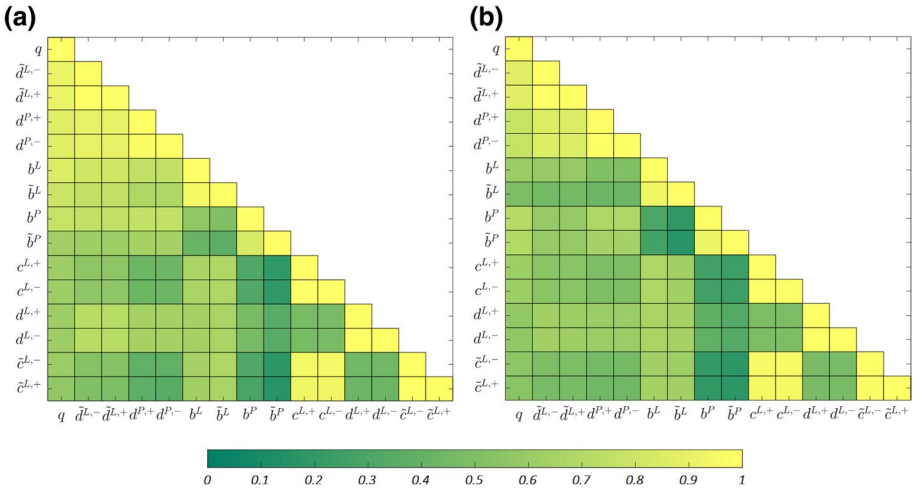

Fig. 7 Illustration of the Pearson correlation coefficient matrices among different variables. **a** The Hague, **b** Amsterdam

on both x and y axis, the sequence of the centrality indicators is arranged in a descending manner based on the correlation between them and $\mathbf{q}$. The sequence for both diagrams is determined by the case of The Hague for the sake of model development.

According to Fig. 7, the in/out-degree centrality indicators in the weighted space-of-infrastructure network ($\tilde{\mathbf{d}}^{\mathbf{L},+/-}$) show the highest positive correlation with $\mathbf{q}$ in both The Hague and Amsterdam tram systems. This is consistent with the visual patterns from Figs. 4 and 5. It is also intuitive to interpret because the amount of passengers that is moved in the network depends on the PT vehicle flows. The following indicators are the in/out-degree centrality in the unweighted space-of-service networks ($\mathbf{d}^{\mathbf{P},+/-}$). Note that these two indicators also show high correlation with the previous ones. They are thus not considered when the in/out-degree centrality in the weighted space-of-infrastructure network are used in the model development.

The group of degree centrality indicators are followed by the betweenness ones. Note that in the case of The Hague (Fig. 7a), the values of betweenness centrality in both of the unweighted and weighted space-of-infrastructure networks ($\mathbf{b}^{\mathbf{L}}$ and $\tilde{\mathbf{b}}^{\mathbf{L}}$) are higher than those in the unweighted and weighted space-of-service networks ($\mathbf{b}^{\mathbf{P}}$ and $\tilde{\mathbf{b}}^{\mathbf{P}}$). This, nevertheless, is opposite in the Amsterdam system (Fig. 7b). In fact, the betweenness centrality in the space-of-infrastructure does not seem to be a good proxy to the passenger flow distribution for the Amsterdam tram network. It performs even worse than the closeness centrality in the space-of-infrastructure. The remaining centrality indicators are presented in the end as they do not show significantly high correlation with $\mathbf{q}$.

Model estimation

The model estimation was performed using MATLAB, with the RE models estimated using the panel data toolbox developed by Álvarez et al. (2017). Note that the robust standard error estimation of the RE models was computed when accounting for heteroscedasticity. Moreover, the variance inflation factor (VIF), which quantifies the severity of collinearity in a regression model, was also computed for the parameters of Model 3 which includes several independent variables.

Three model estimations based on the training dataset from The Hague are presented and discussed in this section, with the detailed results displayed in Table 3. Note that Model 1 is estimated as an ordinary least squares (OLS) model. This is because of the fact that there is no temporal variance in the only independent variable ($\mathbf{b}^{\mathbf{L}}$), and the temporal dimension of the independent variable ($\mathbf{q}$) is correspondingly also canceled by summing up the flows over all periods. This model indicates to what extent it is possible to approximate the global passenger flow distribution using solely topological information without embedding time-dependent service attributes. The other two models, Model 2 and Model 3, are estimated using RE models as explained in "Model development" section because both of them incorporate independent variables pertaining to frequency which is time-dependent. The third model has the highest prediction power also when accounting for the number of parameters included (Adjusted R^2). It includes four centrality indicators: betweenness centrality in the unweighted space-of-infrastructure ($\mathbf{b}^{\mathbf{L}}$), in-degree centrality in the weighted space-of-infrastructure ($\tilde{\mathbf{d}}^{\mathbf{L},+}$), betweenness centrality in the unweighted space-of-service ($\mathbf{b}^{\mathbf{P}}$), and in-closeness centrality in the unweighted space-of-infrastructure ($\mathbf{c}^{\mathbf{L},+}$). The VIF values confirm that all the incorporated independent variables in Model 3 do not exercise significant collinearity since all the values are lower than 10 (Marquardt 1980).

Table 3 Estimation results of the selected models

Independent variables	Model 1 (OLS)			Model 2 (RE)			Model 3 (RE)			
	Coef.	Std.Err	t-stat	Coef.	Rob.Std.Err	z-stat	Coef.	Rob.Std.Err	z-stat	VIF
CONST	− 0.0003	0.0003	− 1.3670	0.0022***	0.0004	5.8880	− 0.0029***	0.0005	− 6.1798	–
b^L	1.0799***	0.0397	27.2241	–	–	–	0.5951***	0.0699	8.5133	3.7495
$\tilde{d}^{L,+}$	–	–	–	0.4849***	0.1168	4.1511	0.1135***	0.0554	2.0491	5.1796
$c^{L,+}$	–	–	–	–	–	–	0.8419**	0.1244	6.7693	1.8418
b^P	–	–	–	–	–	–	0.1140***	0.0089	12.7630	2.5844
Num. Obs.	232			4176			4176			
R^2	0.7632			0.85723			0.89954			
Adj R^2	0.7621			0.85720			0.89944			

***$p < 0.01$; **$p < 0.05$

Table 4 Results of the evaluation metrics for the selected models

Evaluation metrics	The Hague			Amsterdam		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
MAE (pax)	184	260	128	335	305	240
WMAE (pax)	520	841	283	804	715	452
MAPE (%)	77.6	248.3	70.9	71.4	155.7	68.8
WMAPE (%)	42.0	58.7	29.1	55.6	50.6	39.8

Model evaluation

The estimated models are evaluated for the tram networks of The Hague (evaluation dataset) and Amsterdam. The results are summarized in Table 4. Note that the evaluation is performed based on the absolute flows obtained by multiplying the predicted relative flow shares by the total amount of flows in the network. Unsurprisingly, Model 3 largely outperforms Model 1 and Model 2 regardless of the metric used. This suggests that models based on a single centrality indicator that does not incorporate information also from the space-of-service are not able to well capture the correlation. In addition, the discrepancy between weighted and unweighted metrics is striking, implying that significant predictive errors occur to stops with relatively low flows.

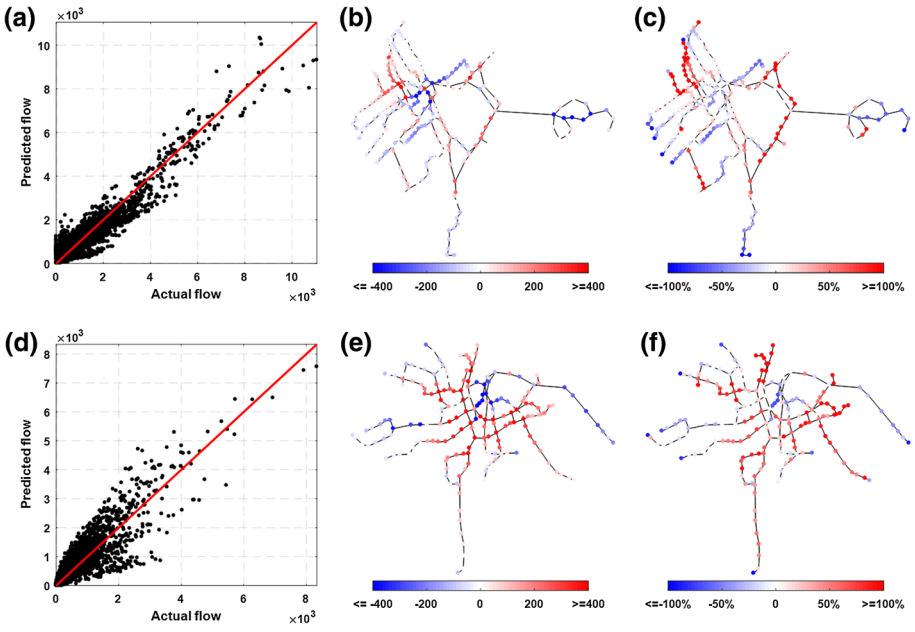

Fig. 8 Illustrations of the evaluation errors for Model 3. **a** Actual flow versus predicted flow plot for The Hague, **b** spatial distribution of the absolute errors for The Hague, **c** spatial distribution of the relative errors for The Hague, **d** actual flow versus predicted flow plot for Amsterdam, **e** Spatial distribution of the absolute errors for Amsterdam, **f** spatial distribution of the relative errors for Amsterdam

Further, we plot the actual versus predicted flows for Model 3 in Fig. 8a, d. It can be observed that Model 3 is indeed well-able to predict passenger flows in both networks. It is also evident that the model performs particularly well when the predictions are made for the same network for which the data has been trained (The Hague).

The spatial distribution of evaluation errors in both absolute and relative terms are also visualized and presented in Fig. 8. Both negative and positive values are considered in the visualizations, corresponding to underestimations (blue) and overestimations (red), respectively. Plots in Fig. 8b, e show absolute error terms, while those in Fig. 8c, f show relative error terms. In the case of The Hague, it can be observed from Fig. 8c that large relative over- or under-estimations occur at stops located further away from the center. However, these relatively large errors in relative terms are small in absolute terms as can be seen in Fig. 8b. In absolute terms, flows at stops in the core of the network tend to be underestimated, while flows along corridors that offer cycles between main parts of the network such as along cross-radial lines are mostly overestimated. Similar overall patterns are observed in the case of Amsterdam, albeit with larger absolute deviations resulting from larger overall demand levels. Hence, flows in the very central core of the network around the central station and the key tourist attractions are underestimated while the flows along the two half-circular infrastructure is overestimated (in both relative and absolute terms for both cases).

Conclusions

This paper presents a pioneering investigation into the relation between passenger flow distribution and network properties in public transport (PT) systems. Differing from the traditional approach that consists of demand estimation and assignment, this study is performed in a reverse engineering fashion by directly examining the relation between the observed flow distribution and network properties that are quantified by centrality indicators in various topological representations of public transport networks (PTNs). This research capitalizes on the capability to measure PT systems using passively collected PT data (e.g., AFC, AVL and GTFS). In addition, concepts and methods adopted from complex network science, including the topological representation of PT infrastructure and service networks and centrality indicators, also play a key role in a sense that the combination of them provides a systematic and concise way to quantify the network properties of PT systems. All the employed centrality indicators are also interpreted in the context of PT systems, which enriches the application of complex network science in the transport research.

The major conclusion drawn from the case study on the tram networks from The Hague and Amsterdam is that the selected network properties can indeed be used to approximate the global passenger flow distribution across the network to a reasonable extent of accuracy using solely regression models. This however does not imply causality as it is likely that supply provision has been designed to correspond to demand patterns and therefore the reflects the interplay between demand and supply distributions. Based on the evidence presented in this paper, several research directions can be further explored in the future. First, more real-world PT systems can be employed in order to further validate the finding. Second, the proposed approach can be instrumental in a range of PT applications. This includes conducting full-scan evaluations of the impact of planned disruption on the redistribution of passenger flows throughout the network, which can serve as a good complement to the prevailing tools, i.e., simulation models,

at a much lower computational cost and with fewer assumptions. Third, the extent to which PT supply is well designed to reflect passenger flow distribution can be considered as a network performance metric for monitoring system performance over time as well as comparing alternative networks.

Acknowledgements The authors thank HTM and Stichting OpenGeo respectively for providing the AFC and AVL data sets of The Hague. The provision of AFC and AVL data sets of Amsterdam by GVB and the help of Dr. Ties Brands during the process are also acknowledged. We acknowledge the support of the SETA project funded by the European Union's Horizon 2020 research and innovation program (Grant No. 688082). An earlier version of this paper was presented on the Conference on Advanced Systems in Public Transport and TransitData 2018 (CASPT2018).

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Akbarzadeh, M., Memarmontazerin, S., Derrible, S., Salehi Reihani, S.F.: The role of travel demand and network centrality on the connectivity and resilience of an urban street system. Transportation (2017). https://doi.org/10.1007/s11116-017-9814-y

Altshuler, Y., Puzis, R., Elovici, Y., Bekhor, S., Pentland, A.: Augmented betweenness centrality for mobility prediction in transportation networks. In: Finding Patterns of Human Behaviors in Network and Mobility Data (NEMO) (2011)

Álvarez, I.C., Barbero, J., Zofío, J.L.: A panel data toolbox for MATLAB. J. Stat. Softw. **76**(6), 1–27 (2017)

Barrat, A., Barthelemy, M., Pastor-Satorras, R., Vespignani, A.: The architecture of complex weighted networks. Proc. Natl. Acad. Sci. U.S.A. **101**(11), 3747–52 (2003)

Bavelas, A.: A mathematical model for group structures. Hum. Organ. **7**(3), 16–30 (1948)

Bavelas, A.: Communication patterns in task-oriented groups. J. Acoust. Soc. Am. **22**(6), 725–730 (1950)

Beauchamp, M.A.: An improved index of centrality. Behav. Sci. **10**(2), 161–163 (1965)

Brandes, U.: A faster algorithm for betweenness centrality. J. Math. Sociol. **25**(2), 163–177 (2001)

Cepeda, M., Cominetti, R., Florian, M.: A frequency-based assignment model for congested transit networks with strict capacity constraints: characterization and computation of equilibria. Transp. Res. Part B Methodol. **40**(6), 437–459 (2006)

Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry **40**(1), 35–41 (1977)

Gao, S., Wang, Y., Gao, Y., Liu, Y.: Understanding urban traffic-flow characteristics: a rethinking of betweenness centrality. Environ. Plan. B Plan. Des. **40**(1), 135–153 (2013)

Gentile, G., Florian, M., Hamdouch, Y., Cats, O., Nuzzolo, A.: The theory of transit assignment: basic modelling frameworks, chap 6. In: Gentile, G., Noekel, K. (eds.) Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems, pp. 287–386. Springer, Berlin (2016)

Guo, Z.: Mind the map! The impact of transit maps on path choice in public transit. Transp. Res. Part A Policy Pract. **45**(7), 625–639 (2011)

Hillier, B., Penn, A., Hanson, J., Grajewski, T., Xu, J.: Natural movement: or, configuration and attraction in urban pedestrian movement. Environ. Plan. B Plan. Des. **20**(1), 29–66 (1993)

Hsiao, C.: Panel data analysis: advantages and challenges. Test **16**(1), 1–22 (2007)

Jiang, B., Liu, C.: Street-based topological representations and analyses for predicting traffic flow in GIS. Int. J. Geogr. Inf. Sci. **23**(9), 1119–1137 (2009)

Kazerani, A., Winter, S.: Can betweenness centrality explain traffic flow? In: 12th AGILE International Conference on Geographic Information Science, Hanover, Germany, pp. 1–9 (2009)

Liu, Y., Bunker, J., Ferreira, L.: Transit users' route-choice modelling in transit assignment: a review. Transp. Rev. **30**(6), 753–769 (2010)

Luo, D., Bonnetain, L., Cats, O., van Lint, H.: Constructing spatiotemporal load profiles of transit vehicles with multiple data sources. Transp. Res. Rec. (2018). https://doi.org/10.1177/0361198118781166
Marquardt, D.W.: Comment: You should standardize the predictor variables in your regression models. J. Am. Stat. Assoc. **75**(369), 87–91 (1980)
Nguyen, S., Pallottino, S.: Equilibrium traffic assignment for large scale transit networks. Eur. J. Oper. Res. **37**(2), 176–186 (1988)
Nuzzolo, A., Russo, F., Crisalli, U.: A doubly dynamic schedule-based assignment model for transit networks. Transp. Sci. **35**(3), 268–285 (2001)
Ortúzar, J.D., Willumsen, L.G.: Modelling Transport, 4th edn. Wiley, New York (2011)
Pelletier, M.-P., Trépanier, M., Morency, C.: Smart card data use in public transit: a literature review. Transp. Res. Part C Emerg. Technol. **19**(4), 557–568 (2011)
Penn, A., Hillier, B., Banister, D., Xu, J.: Configurational modelling of urban movement networks. Environ. Plan. B Plan. Des. **25**(1), 59–84 (1998)
Puzis, R., Altshuler, Y., Elovici, Y., Bekhor, S., Shiftan, Y., Pentland, A.: Augmented betweenness centrality for environmentally aware traffic monitoring in transportation networks. J. Intell. Transp. Syst. Technol. Plan. Oper. **17**(1), 91–105 (2013)
Schmöcker, J.D., Fonzone, A., Shimamoto, H., Kurauchi, F., Bell, M.G.: Frequency-based transit assignment considering seat capacities. Transp. Res. Part B Methodol. **45**(2), 392–408 (2011)
Spiess, H., Florian, M.: Optimal strategies: a new assignment model for transit networks. Transp. Res. Part B Methodol. **23**(2), 83–102 (1989)
Turner, A.: From axial to road-centre lines: a new representation for space syntax and a new model of route choice for transport network analysis. Environ. Plan. B Plan. Des. **34**(3), 539–555 (2007)
van Nes, R., Hamerslag, R., Immers, B.H.: Design of public transport networks. Transp. Res. Rec. **1202**, 74–83 (1988)
van Oort, N., Sparing, D., Brands, T., Goverde, R.M.: Data driven improvements in public transport: the Dutch example. Public Transp. **7**(3), 369–389 (2015)
Vlahogianni, E.I., Park, B.B., Van Lint, J.W.: Big data in transportation and traffic engineering. Transp. Res. Part C Emerg. Technol. **58**, 161 (2015)
von Ferber, C., Holovatch, T., Holovatch, Y., Palchykov, V.: Public transport networks: empirical analysis and modeling. Eur. Phys. J. B **68**, 261–275 (2009)
Wen, T.H., Chin, W.C.B., Lai, P.C.: Understanding the topological characteristics and flow complexity of urban traffic congestion. Physica A Stat. Mech. Appl. **473**, 166–177 (2017)
Ye, P., Wu, B., Fan, W.: Modified betweenness-based measure for prediction of traffic flow on urban roads. Transp. Res. Rec. J. Transp. Res. Board **2563**, 144–150 (2016)
Zhang, Y., Lam, W.H.K., Sumalee, A., Lo, H.K., Tong, C.O.: The multi-class schedule-based transit assignment model under network uncertainties. Public Transp. **2**(1), 69–86 (2010)
Zhao, S., Zhao, P., Cui, Y.: A network centrality measure framework for analyzing urban traffic flow: a case study of Wuhan, China. Physica A Stat. Mech. Appl. **478**, 143–157 (2017)

Ding Luo received his B.Eng. in traffic and transportation and M.Sc. in transport and geoinformation technology from Beijing Jiaotong University and KTH Royal Institute of Technology, respectively. Since 2016, he has been working towards the Ph.D. degree in transportation at Delft University of Technology (TU Delft). His main research interest lies in data-driven analytics and modelling of public transportation systems using various approaches, such as machine learning techniques and complex network science.

Oded Cats is Associate Professor of Passenger Transport Systems at Delft University of Technology (TU Delft). His research develops methods and models of multi-modal metropolitan transport systems by combining advancements from behavioural sciences, operations research and complex network theory. He co-directs the Smart Public Transport Lab at TU Delft where a group of researchers is working extensively with public transport network modelling and passenger demand data analytics.

Hans van Lint received his M.Sc. degree in civil engineering informatics and Ph.D. degree in transportation from Delft University of Technology (TU Delft) in 1997 and 2004, respectively. He was an information analyst and a transport engineer with various organizations. He was appointed as the Anthonie van Leeuwenhoek Full Professor (an honor reserved for only a few young talented scientists and educators) by the Executive Board of TU Delft in 2013. He has co-authored over 60 peer-reviewed journal articles. His

expertise lies on the interface between traffic flow theory and simulation, data analytics, and machine learning techniques. He has co-promoted 8 Ph.D. students and has currently 10 Ph.D. students, five post-doctoral students, and two programmers under supervision in his lab (https://dittlab.tudelft.nl). He serves as an associate editor for IEEE Transactions on Intelligent Transportation Systems, and is active in many international projects and collaborations.