Where shall we sync? Clustering passenger flows to identify urban public transport hubs and their key synchronization priorities

Yap, Menno; Luo, Ding; Cats, Oded; van Oort, Niels; Hoogendoorn, Serge

DOI
10.1016/j.trc.2018.12.013

Publication date
2019

Document Version
Final published version

Published in
Transportation Research Part C: Emerging Technologies

Citation (APA)
Yap, M., Luo, D., Cats, O., van Oort, N., & Hoogendoorn, S. (2019). Where shall we sync? Clustering passenger flows to identify urban public transport hubs and their key synchronization priorities. *Transportation Research Part C: Emerging Technologies*, *98*, 433-448. https://doi.org/10.1016/j.trc.2018.12.013

Transportation Research Part C 98 (2019) 433–448

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Where shall we sync? Clustering passenger flows to identify urban public transport hubs and their key synchronization priorities

Menno Yap*, Ding Luo, Oded Cats, Niels van Oort, Serge Hoogendoorn

Delft University of Technology, Department Transport & Planning, Delft, the Netherlands

ARTICLE INFO

Keywords:
Clustering
Community detection
Hubs
Public transport
Synchronization

ABSTRACT

Minimizing passenger transfer times through public transport (PT) transfer synchronization is important during tactical planning and real-time control. However, there are computational challenges for solving this Timetable Synchronization Problem (TSP) for large, real-world urban PT networks. Hence, in this study we propose a data-driven, passenger-oriented methodology as a preparatory selection stage to reduce problem dimensionality by (1) determining the significant transfer hubs in the network, and (2) identifying subsets of lines within these hubs that need to be prioritized for transfer synchronization. In the first phase of our methodology we determine the spatial boundaries of transfer locations, using a clustering technique based on the passenger transfer flow matrix inferred from smartcard data. After that, a subset of hubs to be prioritized for synchronization is selected. In the second phase, we characterize the transfer patterns within the hubs based on a topological representation. Based on these topological graphs, the line bundles that need to be prioritized within the hubs are further identified using a modularity-based community detection technique. We apply our methodology to a real-world case study, i.e. the PT network of The Hague, the Netherlands. For this case study, our approach allows for prioritizing 70% of all transfers within identified transfer locations while only requiring 0.9% of these transfer locations, thus reducing the complexity of solving the TSP substantially at a relatively low cost. Our method supports public transport operators during timetable design and real-time control in determining where and which lines to prioritize when devising measures for improving transfer experience and synchronization.

1. Introduction

Transfers are an inevitable part of public transport (PT) journeys, since it is not economically viable to directly connect all origin-destination pairs in a network. Transfer locations are however potential weak parts of the total passenger journey experience. Empirical studies show that transfers are perceived one of the most negative components in the public transport journey (e.g. Schakenbos et al., 2016; Van Oort et al., 2016). Improving the transfer experience thus offers potential to improve the attractiveness of public transport and to increase public transport ridership. The passenger transfer experience can be improved by objective and subjective means. Subjective measures focus on improving the waiting time experience at stops during a transfer (e.g. Van Hagen, 2011). On a strategic level, objective measures aim to reduce passenger waiting times by optimizing PT lines (e.g. Gkiotsalitis et al., 2019) and optimizing headways between services (e.g. Gkiotsalitis and Cats, 2018; Varga et al., 2018). On a tactical and operational level, objective measures relate to transfer synchronization, thereby aiming to reduce transfer waiting (and possibly walking) time.

* Corresponding author.
E-mail address: M.D.Yap@TUDelft.nl (M. Yap).

https://doi.org/10.1016/j.trc.2018.12.013
Received 24 March 2018; Received in revised form 19 December 2018; Accepted 20 December 2018
Available online 26 December 2018

Although minimizing passenger transfer waiting time by PT synchronization is important during tactical planning and real-time operations, there are limits for optimization in terms of scalability and complexity. The Timetable Synchronization Problem (TSP), aimed to optimize transfer synchronization in PT networks, has been addressed earlier in many studies in the context of either tactical planning (timetable design) or real-time control. The latter results from stochasticity or disruptions affecting actual vehicle arrival and departure times. In these studies, solving the TSP is usually applied to a relatively small case study network. For example, Lee et al. (2014) consider the impact of synchronizing two lines during tactical planning on service reliability, whereas Gavriilidou and Cats (2018) study real-time synchronization of two tram lines based on passenger data, both for one specific transfer location. Nesheli and Ceder (2015) compare the effects of different real-time control tactics when solving the TSP, applied to a case study network of three bus lines with two transfer locations. Hadas and Ceder (2010) optimize real-time transfer synchronization by simulation of different control strategies, applied to a case study network consisting of one train line, three bus lines and five transfer locations. The abovementioned studies are examples of mathematical programming or control theory approaches applied to the TSP. These studies apply their approach either to a selection of PT lines and transfer locations from the total PT network, or to PT networks constituting small- to medium-sized graphs (e.g. metro networks consisting of a few lines and transfer locations). In contrast, to the best of our knowledge, no studies have been successful in solving the TSP optimization process for large, real-world urban PT networks, often consisting of tens to hundreds of PT lines and transfer locations. Hence, finding or even approximating an optimal solution for the TSP becomes mathematically expensive, if not infeasible.

Solving the TSP is considered NP-hard due to the combinatorial nature of the problem (Desaulniers and Hickman, 2007). For practical problems in larger real-world PT networks, computation time for solving this problem can therefore rise substantially, making it infeasible to solve. Enumeration of all transfer possibilities for a large real-world network would result in a very large number of transfer possibilities, since each transfer location served by $|l|$ unidirectional lines provides $|l| * (|l| - 2)$ transfer possibilities, excluding transfers to the same line. For example, enumerating all transfer possibilities for the urban PT network of The Hague, the Netherlands, provides 1720 transfer possibilities in total between all lines. Solving the TSP for all transfer possibilities of this whole network would become infeasible within reasonable computation times.

To address the computational challenges of solving the TSP for larger, real-world urban PT networks, we propose a methodology for systematically determining where in the network, and for which lines transfer synchronization should be prioritized. Our study thus introduces two steps preceding solving the TSP – identify key priorities (a) where to synchronize, and (b) which lines to synchronize. These two steps are aimed at reducing the combinatorial complexity of the subsequent TSP, and result in a subset of transfer locations and lines where synchronization should be prioritized based on passenger transfer flows. Subsequently, the optimization process to solve the TSP can be applied to this subset, resulting in (an approximation of) optimal synchronization at the most important locations and amongst the most important lines. The novelty of this study lies in problem definition and in using a combination of approaches which have not been used previously for addressing this key planning and operation challenge.

The first step of our methodology is necessary to find a subset of most important transfer locations from the large number of transfer locations a real-world urban PT network consists of, and to determine the spatial demarcation of these urban transfer locations. Given the large number of transfer locations within a real-world urban PT network, the most important transfer locations – defined as hubs – are prioritized during transfer synchronization. Determining the geographical boundaries of transfer locations in high-density urban PT networks is however far from a trivial task. In airline networks for example, the spatial demarcation of a hub is usually unambiguous, given the well-defined physical boundaries of airports and the large distance between airports. The same reasoning applies to train stations being part of an (inter-)regional train network as well. Conversely, the spatial demarcation of transfer locations is less clear for urban PT networks, since many stops are located within walking distance from each other. Locations in the urban network where there is a high flow of transferring passengers (e.g. a bus terminal, train station or shopping area) usually have a large number of tram and bus stops in their surroundings. It is however unclear which of these stops can be considered as one coherent transfer location, and which stops do not make up a part of this transfer location. On the one hand, PT stops which share the same public name, but are located a bit further away from the other PT stops with this name, could constitute part of one large transfer location, not belong to this transfer location, or possibly form a separate, second transfer location, depending on the passenger transfer flows between all different PT stops. On the other hand, given the relatively small distance between different stops in an urban PT network, there can be substantial transfer flows (involving walking) between stops with different public names, which could mean these stops form one transfer location based on passenger flows. It is therefore necessary to develop a methodology to systematically identify the spatial demarcation of urban PT hubs, being the most important transfer locations, without relying only on local knowledge, geographical information or public stop names.

The second step of our methodology uses the identified hubs with their spatial demarcation from the first step, to determine synchronization priorities within each of these hubs. Based on the spatial demarcation, it can be determined which lines within a hub should be considered in the prioritizing process. When the most important transfer locations for transfer synchronization are determined, solving the TSP within a specific hub can still be problematic due to the number of transfer possibilities. For example, a hub with 15PT lines already results in 840 transfer combinations, which makes optimizing coordination between all lines simultaneously computationally challenging. In order to make solving the TSP feasible, there is a need to identify which PT lines in which direction should be considered as one group – in the remainder of this paper coined as one *line bundle* – to be subject to synchronization efforts simultaneously.

We apply cluster analysis as the unsupervised learning technique to identify the hubs and line bundles to synchronize. Machine learning approaches have been applied in a variety of studies related to understanding travel patterns and predicting passenger flows. Examples of studies applying unsupervised learning techniques applied to better understand travel patterns are Agard et al. (2007), Ma et al. (2013, Cats et al. (2015), El Mahrsi et al. (2017) and Luo et al. (2017). Studies performed by Wei and Chen (2012), Ding

Table 1
Indices and sets, variables and parameters.

Indices and sets	
v, V	Index for each node of graph G, set of nodes
e, E	Index for each edge of graph G, set of links
s, S	Index for each public transport stop, set of stops
l, L	Index for each unidirectional public transport line, set of lines
i, I	Index for matrix rows representing origin nodes, set of origin nodes
j, J	Index for matrix columns representing destination nodes, set of destination nodes
t, T	Index for selection of stops identified as transfer location, set of transfer locations
h, H	Index for selection of transfer locations identified as hub, set of hubs
c, C	Line bundle, set of line bundles
Variables	
a_{ij}	Element of weighted adjacency matrix (Section 2.2.2)
d	Scheduled headway
f_{ij}	Passenger transfer flow between i and j
k_t	Passenger transfer flow between all stops that belong to transfer location t
p_{ij}	Element of weighted adjacency matrix of null model (Section 2.2.2)
q	Modularity
w_i	The sum of the weights of the edges attached to node i in the definition of modularity
Parameters	
γ_{max}	Maximum transfer walking distance
ϵ	Maximum distance function value to form a cluster in DBSCAN (Section 2.1.2)
θ	Minimum number of stops required to form a cluster in DBSCAN (Section 2.1.2)
ϑ	Walking speed

et al. (2016) and Li et al. (2017) are exemplary for supervised machine learning applications to better predict passenger flows. The variety of applications of machine learning techniques in public transport demonstrates the potential of using machine learning. Notwithstanding, to the best of our knowledge, machine learning has not yet been applied to identify the spatial boundaries of hubs and line bundles within hubs. This emphasizes the need to develop a new machine learning application to be able to address our research statement to prioritize locations and line bundles for synchronization. In this study, we develop a generic methodology which is data-driven and independent from local network knowledge or a specific network topology by applying unsupervised learning methods. We adopt a passenger perspective, thereby performing our clustering fully based on passenger transfer flow data, rather than using geographical information or incorporating operator constraints in the prioritization. We apply our proposed methodology to the urban PT network of The Hague, the Netherlands as a case study.

The main contributions of this study are therefore (a) the development of a data-driven methodology to identify hubs with their spatial demarcation in urban PT networks; (b) the determination of line bundles within these hubs that need to be prioritized in transfer synchronization, based on a graph representation of transfer patterns and a community detection technique adopted from the field of complex network science. Section 2.1 addresses our methodology for identifying the spatial boundaries of transfer locations and selecting hubs from these transfer locations. Section 2.2 describes the approach to identify line bundles to prioritize in transfer synchronization. In Section 3, the case study is introduced. Section 4 discusses the results of the successive steps of our methodology, followed by conclusions and further research recommendations in Section 5.

2. Methodology

For the remainder of the paper, we introduce indices, sets and variables as presented in Table 1. Furthermore, we introduce the following definitions. A *transfer location* t consists of one or more stops $s \in S$ which are considered one coherent cluster based on passenger transfer flows. Since not all stops $s \in S$ are part of a transfer location, $T \subseteq S$ applies. *Hubs,* H*,* are the most important transfer locations amongst T based on the passenger transfer flows between stops, so that $H \subseteq T$ applies. Each public transport line $l \in L$ represents a route with a unique public line number as communicated to passengers in a certain direction and is therefore unidirectional. Each cluster of unidirectional public transport lines which should be considered as one group simultaneously during synchronization at a hub $h \in H$ is defined as a *line bundle* $c \in C$, with $c = \{l_1^c, l_2^c, \cdots, l_n^c\}$.

Fig. 1 presents a flowchart with all steps of the proposed methodology. It shows the methodology used to identify hubs and their spatial demarcation to prioritize for synchronization (Section 2.1), and the methodology used to identify the line bundles to synchronize simultaneously within each identified hub (Section 2.2).

2.1. Identification of transfer location priorities for synchronization

2.1.1. Infer stop-to-stop transfer flow matrix

The first step of our methodology is to infer the stop-to-stop transfer flow matrix, as visualized by the first row of phase 1 in Fig. 1. As input for our study we use passenger data from Automated Fare Collection (AFC) systems and PT vehicle position data obtained

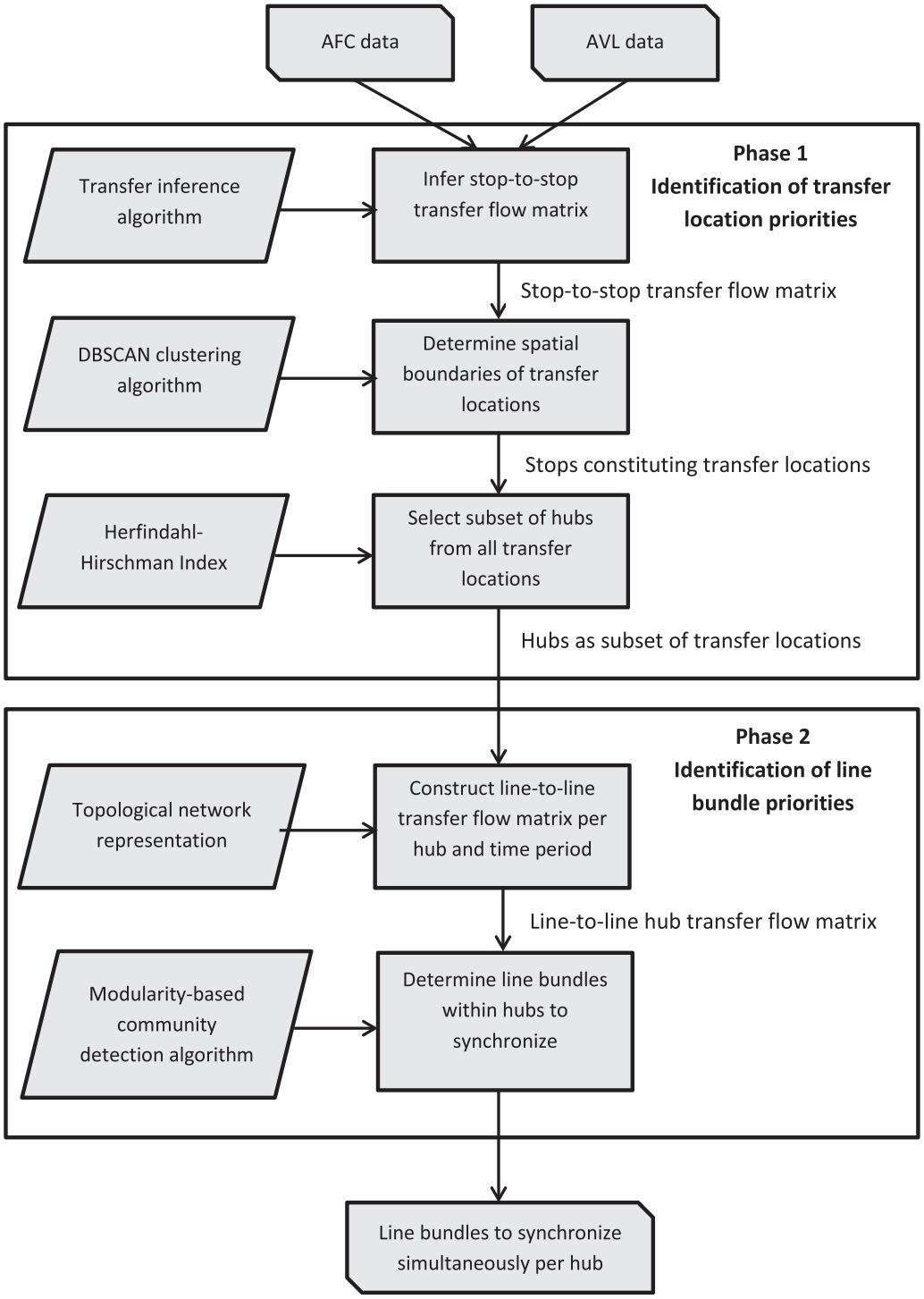


Fig. 1. Flow chart of the proposed methodology.

from Automated Vehicle Location (AVL) systems for urban tram and bus services. Each transaction from AFC systems contains at least the passenger smart card id, tap-in time, tap-in stop, and the vehicle line and trip-id of the run a passenger boarded for each journey leg separately. In some cases – such as Seoul, Queensland and the Netherlands – also the passenger tap-out time and tap-out stop are registered. In case of entry-only AFC systems, the alighting location can be inferred using different destination inference algorithms, of which a trip-chaining algorithm is most commonly applied (e.g. Trépanier et al., 2007; Zhao et al., 2007; Nunes et al., 2016; Munizaga and Palma, 2012). This results in AFC transactions with information about the passenger tap-in and tap-out time and

Table 2
Illustration of format AFC data.

Tap-in date and time	Tap-in stop-ID	Tap-in line	Tap-out date and time	Tap-out stop-ID	Trip-ID	Vehicle ID	Smart-card ID
4-3-2018 11:42:37	35,309	6	4-3-2018 12:03:19	34,997	3423	3050	81,675,688
4-3-2018 12:15:57	30,091	18	4-3-2018 12:23:04	32,857	6545	187	81,675,688

location used as input for our study (see Table 2 as illustrative data format). Each row of the AVL data contains information about the scheduled and realized arrival time and departure time of each PT run with corresponding trip-id at each stop (see Table 3 for illustration purposes). Besides, the coordinates of each PT stop of the considered urban PT network are used as input. Each separate platform has a unique stop-id and unique coordinates.

Since AFC data contains passenger transactions per journey leg for urban tram and bus services, we apply a transfer inference algorithm to determine which transactions constitute one passenger journey. In this study we apply the transfer inference algorithm proposed by Yap et al. (2017), which elaborates on previous work by Gordon et al. (2013). This algorithm distinguishes transfers from final destinations using assumptions on passenger behaviour during both undisrupted and disrupted scenarios. Using this transfer inference algorithm has therefore the advantage that no a priori data cleaning or data classification as disrupted or undisrupted is needed, making the need to demarcate the time the passenger impact of a disruption starts and ends obsolete.

To execute this algorithm, we fuse the AFC and AVL datasets based on the trip-id variable both data systems have in common. For urban tram and bus networks with 100% smart card penetration rate, in-vehicle devices and registered or inferred tap-out location, the stop-to-stop vehicle occupancies for each trip-id can be directly obtained from fusion of AFC and AVL data. In case the smart card penetration rate is not 100%, the obtained occupancies should be increased by the percentage non-card users obtained from alternative data sources, such as passenger counts. Below we shortly describe the applied transfer inference algorithm. For a more extensive explanation we refer the reader to Yap et al. (2017). An alighting is considered a transfer if it satisfies the following three conditions:

- Temporal condition: an alighting passengers boards the first vehicle of the subsequent line after the first reasonable passenger arrival time at the next boarding stop – based on the transfer walking distance and assumed walking speed ϑ – where the vehicle occupancy does not exceed the norm capacity;
- Spatial condition: the next boarding location does not exceed a maximum transfer walking distance threshold γ_{max} from the previous alighting stop;
- Line-based condition: the next boarding line is not the same line as previously alighted from, or the next boarding line is the first run of this same line after the alighted run, in order to incorporate the impact of possible rescheduling measures during disruptions, such as short-turning or deadheading.

After applying this transfer inference, a stop-to-stop transfer flow matrix can be constructed from the alighting stop and next boarding stop for each alighting which is considered a transfer.

2.1.2. Determine the spatial boundaries of transfer locations

In the next step (visualized by the second row of phase 1 in Fig. 1) we determine the spatial demarcation of transfer locations by applying a clustering algorithm. This entails determining which urban PT stop-ids form a coherent cluster of stops between which passenger transfer flows occur. To this end, we apply a density-based clustering technique. This data-driven approach for determining the spatial boundaries of transfer locations implies that the number of stops each cluster is composed of is not known a priori. The total number of clusters is thus also not pre-determined, making k-means/k-medoid clustering not suitable for this purpose. Moreover, our clustering should also not be collectively exhaustive: the considered PT network includes transfer locations consisting of one stop-id only based on passenger transfer flows, such as one platform being served my multiple PT lines with transferring passengers between these lines. Consequently, clustering only needs to be applied for locations where transfers occur between multiple PT stops (multiple platforms). This requirement makes hierarchical agglomerative clustering, a collectively exhaustive clustering technique, not suitable for purpose. A density-based clustering technique which allows for partial clustering without a pre-defined number of clusters (e.g. DBSCAN, OPTICS) fulfills all of the aforementioned requirements (Tan et al., 2004). We apply DBSCAN, the most commonly applied density-based clustering technique. For an in-depth explanation of the algorithm of DBSCAN we refer to Ester et al. (1996).

In traditional geographical applications of DBSCAN, a geographical measure – such as the Euclidean or geodesic distance between nodes – is applied as distance measure for clustering. This means that the closer two nodes are geographically positioned to each

Table 3
Illustration of format AVL data.

Stop-ID	Trip-ID	Scheduled arrival time	Realized arrival time	Scheduled departure time	Realized departure time
1119	4464	2017-01-06 19:22:35	2017-01-06 19:23:25	2017-01-06 19:22:35	2017-01-06 19:23:49
1119	4465	2017-01-06 18:23:48	2017-01-06 18:26:26	2017-01-06 18:23:48	2017-01-06 18:26:44

other, the more likely that these nodes are being grouped into the same cluster. A larger distance value thus indicates a lower clustering likelihood. However, to identify synchronization priorities purely based on passenger demand, we cluster in our study purely based on passenger flows rather than geographical distances. We define $f_{s_i s_j}$ as the obtained transfer flow between origin PT stop i and destination PT stop j, where each urban PT stop $s \in S$ indicates a unique stop code in the considered urban PT network and $|S|$indicates the number of stops (see Table 1). To create a symmetrical distance matrix, we first sum transfer flows $f_{s_i s_j}$ and $f_{s_j s_i}$ to $g_{s_i s_j}$ (Eq. (1)). In this context, contrary to traditional geographical distance measures in DBSCAN, a higher transfer flow between stops thus *increases* the likelihood of these stops being clustered together. This means that in our case the regular distance measure cannot be applied for DBSCAN. Values $g_{s_i s_j}$ are therefore transformed into $g'_{s_i s_j}$, such that a higher value decreases the clustering likelihood. By subtracting $g_{s_i s_j}$ from the maximum value $\max_{s_i s_j \in S} g_{s_i s_j}$, all values remain non-negative (Eq. (2)). Clustering based on distance measure $g'_{s_i s_j}$ entails stops being clustered if there are strong transfer flows between them. This implies that stops which are geographically close to each other do not necessarily have to be clustered together, if there is no substantial transfer flow between these stops.

$$g_{s_i s_j} = f_{s_i s_j} + f_{s_j s_i} \quad \forall\, s_i, s_j \in S \tag{1}$$

$$g'_{s_i s_j} = \max_{s_i s_j \in S} g_{s_i s_j} - g_{s_i s_j} \quad \forall\, s_i, s_j \in S \tag{2}$$

Two parameters θ and ϵ need to be specified in the DBSCAN algorithm. θ indicates the minimum number of stops required to form a cluster. We set this value to one, meaning that the inclusion of at least one other stop (two stops in total) is required for the minimum cluster size. Since we perform a partial clustering where not all stops need to be clustered, each cluster is composed of at least two stops. ϵ indicates the maximum distance function value required when forming a cluster. For our formulated distance measure $g'_{s_i s_j}$, this means setting a minimum requirement for the transfer flow between stops that form a cluster. In line with the recommendation by Ester et al. (1996), we plot the θ-distance graph for different values of ϵ to determine the knee in the plot to identify a suitable parameter value. This total distance measure is calculated by averaging the minimum distance value of all identified clusters for a given value of ϵ (Eq. (3)), where $t \in T$ is a transfer location resulting from DBSCAN and T is the set of all identified transfer locations (see Table 1). After applying DBSCAN it can be determined which PT stops constitute a certain transfer location, which shows the spatial boundaries of each transfer location.

$$dist = \frac{\sum_{t \in T} \min_{s_i s_j \in t} g'_{s_i s_j}}{|T|} \tag{3}$$

2.1.3. Select subset of hubs from all transfer locations

Not all identified transfer locations are considered a hub, given substantial differences in magnitude of transfer flows within each transfer location. The aim of this step is to identify the set of hubs H, which is a subset from all transfer locations $H \subseteq T$. The transfer locations with the largest transfer flows between stops are considered a hub (as visualized in the third row of phase 1 in Fig. 1). We apply a method used to determine hubs in airline networks based on Costa et al. (2010), since no comparable method has yet been applied to identify hubs in urban PT networks. First, the total intra-transfer location transfer flow k_t between all stops is calculated for each identified transfer location t (Eq. (4)). Based on the work of Costa et al. (2010), we apply the Herfindahl- Hirschman Index (HHI) to calculate the market concentration based on the market share of each transfer location $t \in T$ in terms of passenger transfer flow as a share of the total transfer flow in the considered PT network. The integer number of hubs $|H|$ can then be determined based on n, which equals the inverse of the HHI (Eq. (5)). In Eq. (5), n can be an integer or a decimal value. When all transfer locations are ranked in decreasing order based on k_t, only the top $|T| < n$ clusters are considered hubs, based on which the set of hubs H is determined. The steps of phase 1 of our proposed methodology for identifying hubs are illustrated in Fig. 2.

$$k_t = \sum_{s_i \in S_t} \sum_{s_j \in S_t} f_{s_i s_j} \quad \forall\, t \in T \tag{4}$$

$$n = \frac{1}{\sum_{t \in T} \left[\left(\frac{k_t}{\sum_{t \in T} k_t} \right)^2 \right]} \tag{5}$$

2.2. Identification of line bundle priorities for synchronization

A complex-network theoretic approach is developed to identify line bundles – in literature sometimes referred to as cliques or communities – within a hub to prioritize simultaneously for synchronization. The proposed approach consists of two steps: (1) establishing the topological representation of the transfer pattern among unidirectional lines within the identified hubs from phase 1; (2) identifying the line bundles within each hub using a community detection technique. This approach provides a data-driven solution that is automated, intuitive and scalable. The details of these two components are presented in the following sections.

2.2.1. Topological representation of the transfer pattern within hubs

In this step of our methodology (as visualized in the first row of phase 2 in Fig. 1) the transfer topology within an identified hub is represented as a directed graph $G = (V, E)$. This representation is inspired by the C-space topological representation of PT networks

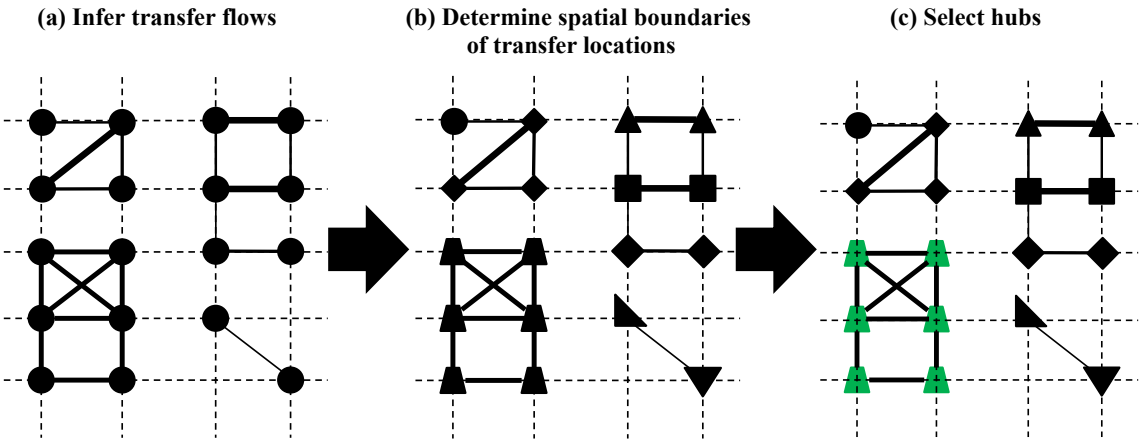


Fig. 2. Illustration of phase 1 of our methodology for a single hypothetical PT network. Nodes reflect stops, and the width of the edge represents the strength of passenger transfer flow. Based on the transfer flows between stops (a), DBSCAN identifies which stops form a cluster of transfer stops (indicated in (b) by the same shape). Applying the HHI identifies hubs as most important transfer locations in (c) (marked green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

where individual lines are represented as nodes and are connected via an edge only if they share common transfer stops (see von Ferber et al. (2009) for a detailed description of the C-space representation of PT networks). In our case, each node $v \in V$ corresponds to a PT line in a certain direction $l \in L$, whereas each edge $e \in E$ represents the observed transfer activity between two lines in certain directions within an identified hub. An illustration of such topological representation is sketched in Fig. 3. Furthermore, the graph G is represented by a weighted adjacency matrix A where a_{ij} denotes the weight of the edge between i and j. We consider two different types of weights in this study, namely the passenger transfer flow (Eq. (6)) and the passenger transfer waiting time (Eq. (7)). The objective of applying two different link weights is to compare clustering results between the case where only passenger transfer flows are considered, and the case when the expected transfer time is incorporated as well.

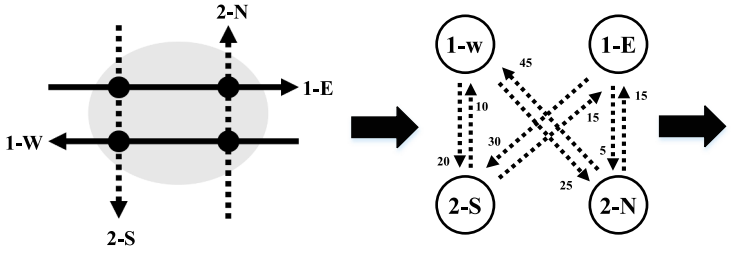
The first type of link weight corresponds to the number of passengers transferring between two lines in a certain direction at hub $h \in H$, which is defined using Eq. (6). In this equation $f^h_{l_i l_j}$ denotes the observed transfer flow from unidirectional line l_i to line l_j within a hub. Values for $f^h_{l_i l_j}$ are derived from the stop-to-stop transfer flow matrix obtained by fusion of AFC and AVL data (as explained in Sections 2.1.1 and 2.1.2), for which transfer flows between stops which are part of the same hub between lines l_i and l_j are summed. The second type of link weight relates to the total passenger transfer waiting time between two unidirectional lines, which is calculated using Eq. (7). Variables $f^h_{l_i l_j}$ and $d^h_{l_j}$, respectively, denote the observed transfer flow between line i and j as calculated using Eq. (6), and the planned headway of line l_j at hub h. Given our focus on high frequent urban PT networks, the assumption of random passenger arrivals at the stop can be justified. In future work a more advanced passenger arrival and waiting time distribution, as proposed by Ingvardson et al. (2018), can potentially be applied to our method.

$$a^h_{ij} = f^h_{l_i l_j} \tag{6}$$

$$a^h_{ij} = \frac{d^h_{l_j} * f^h_{l_i l_j}}{2} \tag{7}$$

2.2.2. Determine hub line bundles for synchronization

Based on the graph representation with the link weighted by transfer flow or transfer waiting time, a community detection



	1-W	1-E	2-S	2-N
1-W	0	0	20	25
1-E	0	0	30	5
2-S	10	15	0	0
2-N	45	15	0	0

Fig. 3. Illustration of the topological representation for the hub transfer pattern. The original layout of the identified hub (shaded area) is presented on the left with four directed lines marked, i.e. 1-E, 1-W, 2-S and 2-N. The transfer pattern is then represented as a graph (middle). The weighted adjacency matrix is displayed on the right.

technique from the field of complex network science is applied to identify line bundles within a hub (see the second row of phase 2 in Fig. 1). In essence, the problem that community detection intends to address is to partition a network into communities of densely connected nodes, with the nodes belonging to different communities being only sparsely connected. Such technique is for example applied by Yildirimoglu and Kim (2018) to identify a community structure in urban mobility networks. In our application, line bundles will thus become the partitioning result based on passenger transfer flows or transfer waiting time used as the link weight, in which intra-community transfer connections are maximized while inter-community values are minimized.

Given our aforementioned objective, an optimization-based method called the Louvain method is adopted to identify hub line bundles. Proposed by Blondel et al. (2008), the Louvain method is a heuristic method based on modularity optimization. As a class of community detection method that has received the greatest attention from researchers, the optimization technique aims at finding an extremum – usually the maximum – of a function indicating the quality of a clustering, over the space of all clustering possibilities (Fortunato and Hric, 2016). The most popular quality function is the *modularity* proposed by Newman and Girvan (2004), which estimates the quality of a partition of the network in communities. The essential idea of this measure is to reveal how non-random the network structure is by comparing the actual structure and its randomization where network communities are destroyed. The value of modularity q varies between -1 and 1, which measures the density of links inside communities as opposed to links between communities. Its general expression is formulated using Eq. (8).

$$q = \frac{1}{2|E|}\sum_{ij}(a_{ij} - p_{ij})\delta_{c_i,c_j} \tag{8}$$

In this equation $|E|$ represents the number of edges of the graph. The summation runs over all pairs of nodes i and j, in which a_{ij} and p_{ij} denote the element of the adjacency matrix and the randomized null model term, respectively. Derived by randomizing the original graph, the term p_{ij} indicates the average adjacency matrix of an ensemble of networks to preserve some of its features. c_i indicates the community to which node $i$is assigned. The Kronecker delta function δ_{c_i,c_j} is defined using Eq. (9).

$$\delta_{c_i,c_j} = \begin{cases} 1, & if \quad c_i = c_j \\ 0, & otherwise \end{cases} \tag{9}$$

The regular modularity function of the Louvain method only uses the adjacency matrix to perform the community detection. However, for our study purpose we consider not only the question whether two nodes are connected in the graph or not, but also how many passengers are transferring between these two nodes. To incorporate passenger transfer flows or transfer waiting time in the community detection, the modularity function is adjusted using these weighted links as shown by Eq. (10), where $w_i = \sum_j a_{ij}$ denotes the sum of the weights of the edges attached to node i (Newman, 2004).

$$q = \frac{1}{2|E|}\sum_{ij}(a_{ij} - \frac{w_i w_j}{2|E|})\delta_{c_i,c_j} \tag{10}$$

The modularity measures essentially how different the original graph is from a randomized graph. The Louvain method is adopted because it has been recognized as one of the best-performing clustering algorithms after a comparative evaluation (Lancichinetti and Fortunato, 2009). The Louvain method has several advantages. First, the algorithm is intuitive and easy to implement. Second, the outcome is unsupervised and computationally light, which requires the link label matrix as the only input. The essence of this method is a greedy optimization of q in a hierarchical manner. It assigns each node to the community of their neighbours that can yield the largest q, and thus creates a smaller weighted super-network whose nodes are the clusters already found. Therefore, partitions found on this super-network consist of clusters that contain previous ones as well, resulting in a higher hierarchical level of clustering. This procedure is not stopped until the largest possible modularity value is reached. A visualization of the steps of this algorithm is presented in Fig. 4.

3. Case study

We apply our methodology to the urban PT network of the city of The Hague, the Netherlands, operated by HTM (Fig. 5). The Hague is one of the main cities of the so-called Randstad, the most important economic area in the western part of the Netherlands and has more than 500,000 inhabitants. The urban PT network of The Hague consists of 12 tram lines and 10 urban bus lines at the time of consideration (November 2015). We only consider the urban PT network, meaning that services on the (inter)regional train network level and regional bus services are not incorporated. On an average working day there are more than 300,000 AFC transactions within the case study network.

As initial input we use all AFC transactions on the case study network for all 20 working days between November 2nd and November 29th, 2015, together with all AVL transactions for this period. For both the identification of hubs and of line bundles within hubs, we only incorporate AFC data for the AM peak and PM peak of working days in which no large disruptions occurred during the AM and PM peak. Since disruptions can result in adjusted passenger route choice including transfer line choice, this can introduce bias when determining the key hubs and line bundles. Based on disruption log-data provided by the operator of this network, we removed data from 10 working days. Our resulting dataset thus contains AFC data from 10 working days of the above-mentioned period (2 Mondays, 1 Tuesday, 3 Wednesdays, 2 Thursdays and 2 Fridays). No destination inference is needed for our case study, given that passengers are required to tap-in and tap-out using devices located within urban trams and buses in The Hague. Incomplete AFC records (1.3%) and AFC records where system errors occurred ($< 0.4\%$) have been removed. The resulting dataset

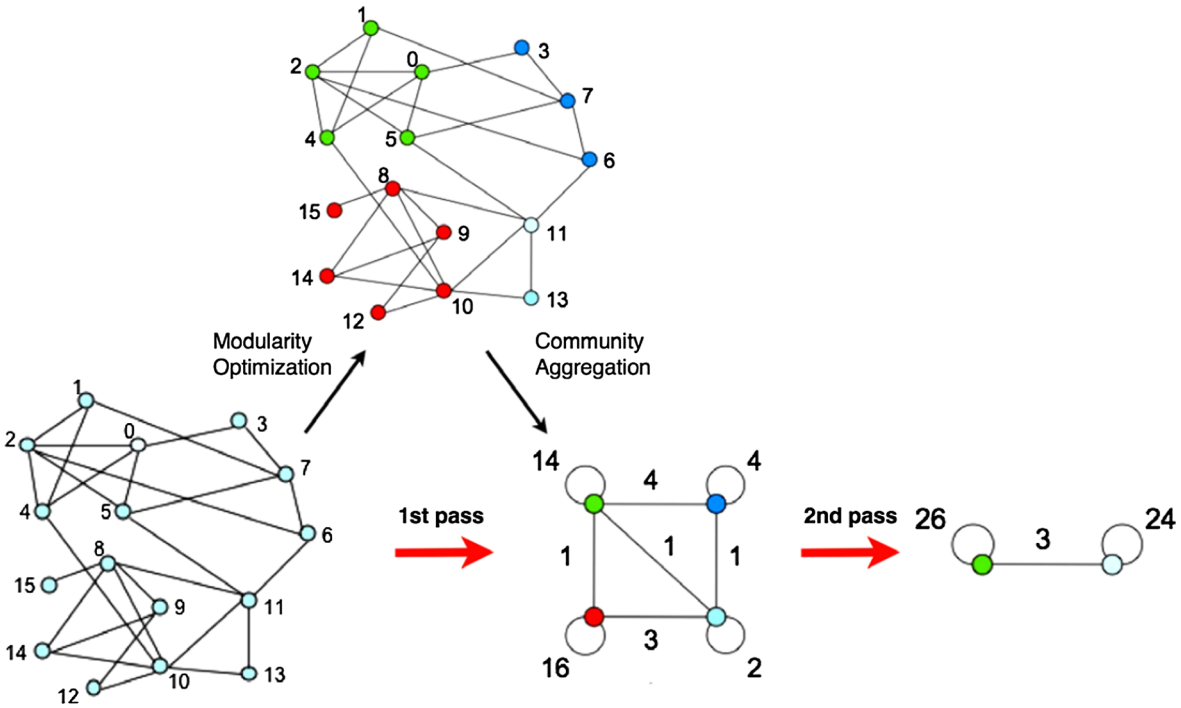


Fig. 4. Visualization of the steps of the Louvain method. Each pass consists of two phases. In the first phase, the modularity is optimized by allowing only local changes of communities; in the second one, the communities found are aggregated in order to build a new network of communities. The passes are repeated iteratively until no increase in modularity is possible (Blondel et al., 2008). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

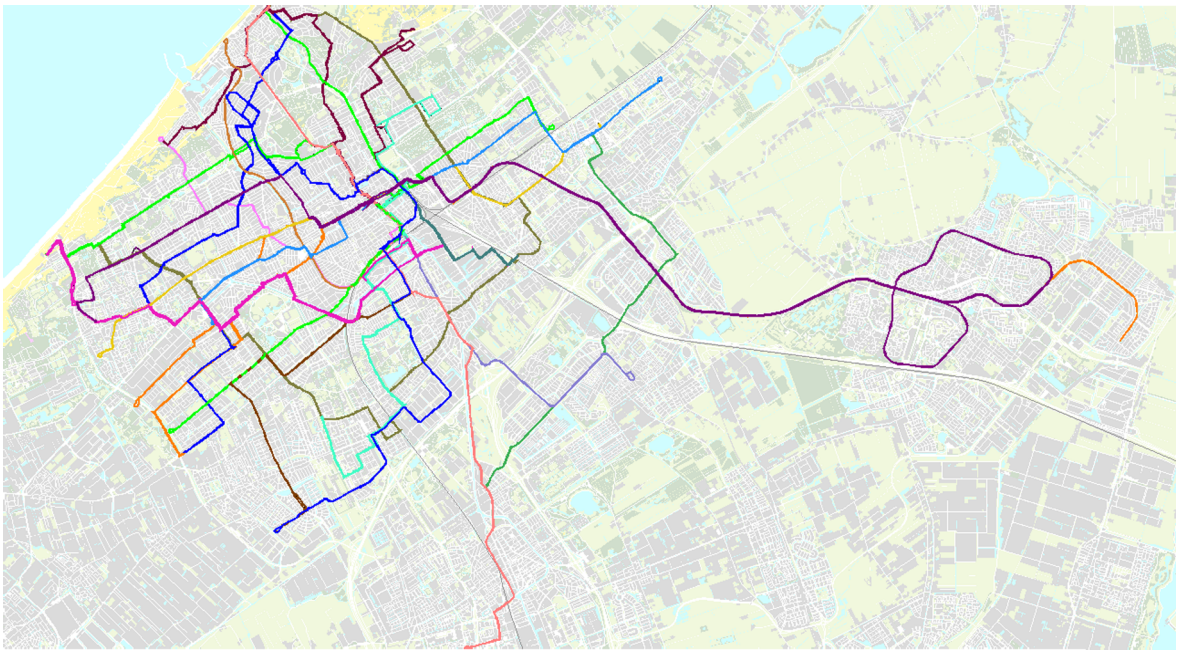


Fig. 5. Overview of urban tram and bus services of the case study network in The Hague.

contains a total of 3.04 million AFC transactions to which we applied our transfer inference algorithm. Since we only focus on AM and PM peak journeys, transactions part of journeys with a starting time outside the interval 07:00–09:00 or 16:00–18:00 have been removed after this step from our dataset, which resulted in 1.1 million AFC transactions remaining for our analysis. The steps we applied in the data cleaning process are summarized in Table 4.

Table 4
Data cleaning process.

Data cleaning process	# Transactions	% Transactions
Initial AFC transactions for 10 working days	3,086,453	100%
Incomplete AFC transactions (missing tap-out)	−40,195	−1.30%
System error transactions	−11,162	−0.36%
AFC transactions part of journey started outside AM or PM	−1,930,711	−62.6%
Complete AFC transactions part of journey stared in AM or PM	1,104,385	35.8%

In phase 1 of our proposed methodology, we identify the set of hubs H and their spatial demarcation for our case study network. For this phase, we use the combined AM and PM transfer flows over the 10 working days, since the definition of a hub is considered independent from the period of the day. In phase 2 of our methodology, we determine line bundles to prioritize in synchronization for each identified hub. Since passenger (transfer) flows can differ substantially over the day, synchronization priorities might differ as well. Therefore, we apply the second phase of our methodology for each time period separately. We compare the clustering results in phase 2 when using transfer flow or transfer waiting time as link weights, which results in four cases per hub. Table 5 shows an overview of all cases considered in phase 2 of our study.

4. Results and discussion

This chapter discusses the results and implications of the hub identification phase (Section 4.1) and line bundle identification phase (Section 4.2) of our proposed methodology.

4.1. Hub identification

We applied the proposed transfer inference algorithm to all 1,104,385 AFC transactions using the Euclidean distance between stops as transfer walking distance, the 2.5th percentile walking speed ϑ from a normal distributed walking speed function $N(1.34, 0.34)$ based on Hänseler et al. (2016), and γ_{max} of 400 Euclidean metres. The value for the maximum transfer walking distance threshold γ_{max} is obtained from Yap et al. (2017), where this showed to be the optimal value when validating the destination inference algorithm applied in that study. This results in 150,792 alighting transactions which are considered a transfer. The sum of the obtained stop-to-stop transfer flow matrix for the AM and PM period together for the considered 10 working days thus equals 150,792. We performed a sensitivity analysis with respect to γ_{max}. Increasing this value by 50% (600 Euclidean metres) and 100% (800 Euclidean metres) reduces the number of identified separate journeys by merely 1% and 2%, respectively. We thus conclude that our results are robust to different values of γ_{max}. While using on-street distances rather than Euclidean distances can further improve the accuracy of the transfer inference algorithm, it might require use of data sources and software which are not always easily accessible. Using on-street distances might be particularly relevant if the methodology would be applied to a network with large variations in street lay-out, such as an application which considers both urban and inter-urban PT networks rather than urban PT networks only.

All 150,792 classified transfers occur between 754 different stop-ids, resulting in a 754×754 transfer flow matrix. We performed DBSCAN to identify the geographical boundaries of clusters of transfer locations and set θ equal to 1 (see Section 2.1.2). The θ-distance plot is shown as function of average distance *dist* (Fig. 6) and as function of the number of identified clusters (Fig. 7). These functions are used to determine the optimal parameter value of ϵ, which reflects a minimum requirement for the transfer flow between stops that form a cluster. Since data from 10 working days are included, we use a step size of 50 for values of ϵ in the 1-distance plot to maintain a meaningful interpretation of this value (using daily integer transfer flows with step size 5, starting with $\varepsilon = \max_{s_i s_j \in S} g_{s_i s_j} = 5516$). As can be observed from Fig. 6, the shape of the graphs is opposite of the regular shape of the θ-distance plot when applying DBSCAN. This is because in our application a higher transfer flow between stops increases the probability of being clustered together, in contrast to traditional distance measures where a larger value entails a smaller clustering probability. In both Figs. 6 and 7 the knee in the graph can be observed for $\varepsilon = 5166$ and 23 identified clusters of transfer locations. From all 754 transfer stops in the case study network, 11% (81 stops) is part of a cluster of at least two stops. The average cluster size equals 3.54; the median cluster size is 3. The largest cluster consists of 11 stops. The other 673 transfer stops are not clustered by DBSCAN and form a

Table 5
Overview of cases in study phase 2 for line bundle identification.

	Link weight: passenger transfer flow		Link weight: passenger transfer waiting time	
Hub	AM	PM	AM	PM
h_1	Case 1A	Case 1B	Case 1C	Case 1D
h_2	Case 2A	Case 2B	Case 2C	Case 2D
$h_{\|n\|}$	Case nA	Case nB	Case nC	Case nD

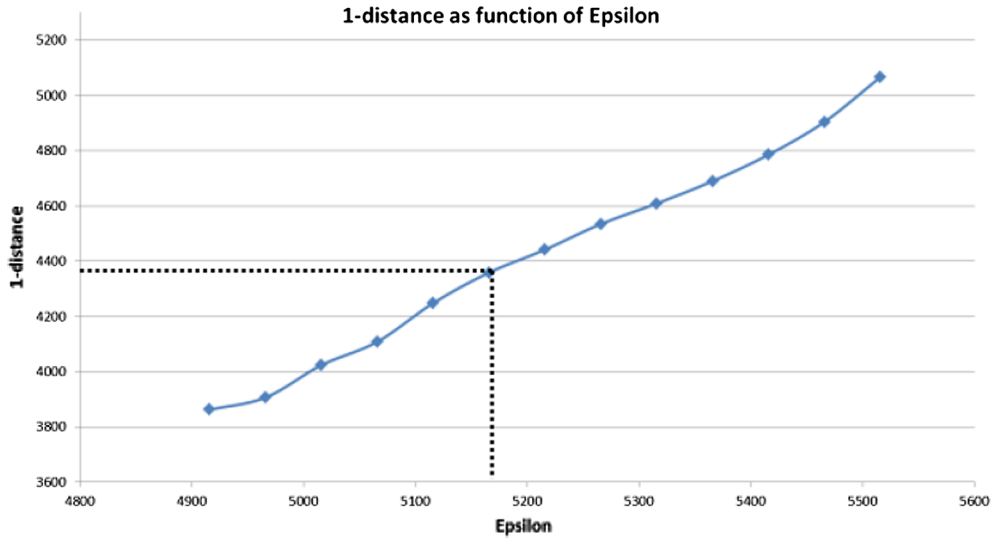


Fig. 6. 1-distance plot as function of ϵ.

Fig. 7. 1-distance plot as function of the number of identified clusters.

transfer location consisting of one stop only. This means that a total of 696 transfer locations with their geographical boundaries are identified for the case study network. Fig. 8 presents all identified transfer locations, highlighting the 23 clusters.

After applying the HHI, 6 transfer locations with the highest number of intra-cluster transfer flows are identified as hub: Central Station (hub 1), Centre (hub 2), Station Hollands Spoor (hub 3), Leidschenveen (hub 4), Leyenburg hospital (hub 5) and Brouwersgracht (hub 6). The identified hubs 1, 2, 3 and 5 are locations where a large number of tram and bus lines intersect near a train station (1 and 3), the city centre (2) and a hospital (5). Hubs 4 and 6 are served by fewer lines: these hubs are mainly characterized by large transfer flows between a corridor of high-frequency tram lines and one intersecting tram line (4) or bus line (6). The stops constituting these hubs are shown in Fig. 8. In the box of Fig. 8, hub 3 (Station Hollands Spoor) is shown in greater detail. This hub illustrates a key difference between the stops which would be considered as one hub purely based on the geographical location or public name of the stop, and the stops found to constitute a hub based on passenger flows. At the north-side of this station, there are several tram stops located relatively close to each other (star-shaped blue nodes indicated by 'A'), whereas a few bus stops of this station are located at the south-side of the station, about 3 min walking from each other (star-shaped blue nodes indicated by 'B'). Besides, there are stops belonging to another public stop name 'Rijswijkseplein', about 5 min walking from the tram stops of the station itself. Our clustering results show that from a passenger perspective some of the stops of Rijswijkseplein are part of one large hub (star-shaped blue nodes indicated by 'C'), whereas they would be considered a separate transfer location if clustered based on geographical location or public stop name. Some other stops of Rijswijkseplein form a separate transfer location but are not part of hub 3 (circular red nodes indicated by 'D').

Fig. 9 (left) shows the cumulative distribution function (CDF) of all transferring passengers k_t for the 696 transfer locations of the case study network. In Fig. 9 (right) the Lorentz-curve is plotted, where the realized transfer flow distribution over the transfer

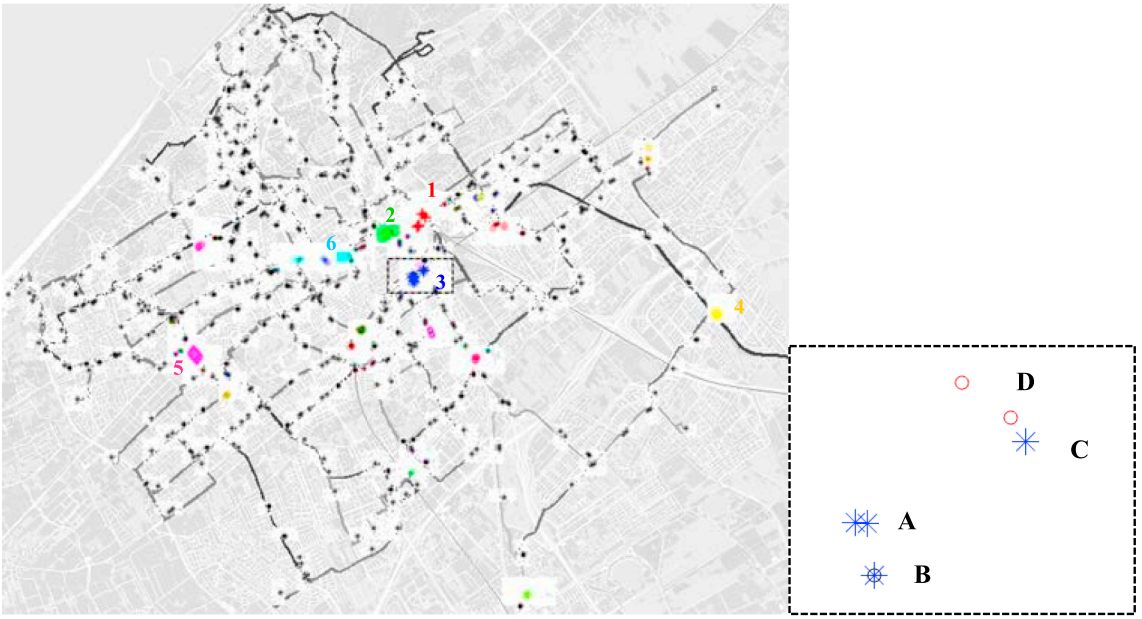


Fig. 8. Locations of identified transfer locations with 23 transfer clusters (all groups of coloured stops) and the 6 selected hubs (coloured stops indicated by numbers 1–6). The box on the right-hand side of the figure zooms into the stops around hub 3 (station Hollands Spoor): blue stops form one hub; red stops (with public name 'Rijswijkseplein') form another transfer location, but are not part of the hub. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

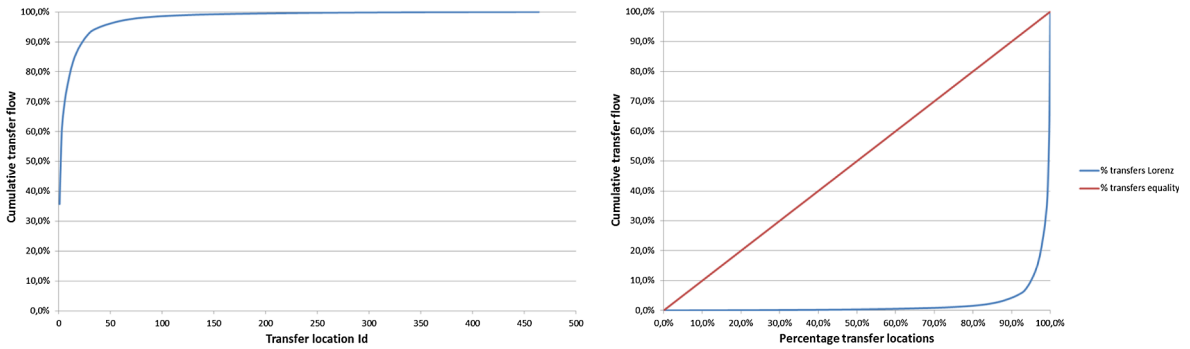


Fig. 9. CDF and Lorentz-curve for transfer flow distribution over the identified transfer locations. 100% equals the total within-cluster transfer flow; between-cluster transfer flows are not incorporated in the figures. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

locations is contrasted with the hypothetical scenario if transfers would be equally distributed over all transfer locations. It can clearly be observed that transfer patterns are not uniformly distributed over all locations. In contrast, a sharp concentration of transfers at a few locations can be observed. The calculated Gini-coefficient of 95.7% from Fig. 9 (right) confirms the unequal spatial distribution of passenger transfers over the urban PT network. The total intra-hub transfer flows, summed over the 6 identified hubs, represent 70.1% of the transfers k_t of all transfer locations, whereas 78.4% of all network transfers are within one of the 696 transfer locations. Our results show that more than 70% of the transfers within all transfer locations can be captured in the optimization, while only 6 out of 696 (0.9%) identified transfer locations needs to be incorporated. Given our aim not to be exhaustive, these results show that the complexity of solving the TSP can be reduced substantially, against relatively limited costs. As can be seen from Fig. 9 (left), considering the top 6 transfer locations is an effective means to capture a large part of the total network transfer flow. Given the unequal spatial transfer distribution, the efficiency of each additional transfer location added to the optimization problem will decrease: the complexity of the TSP increases, whereas the number of additional captured transfers only decreases compared to the prior ranked transfer location.

4.2. Line bundle identification

Table 6 summarizes the statistics for the line bundle identification phase of our methodology. Given that 6 hubs were identified,

Table 6
Summary statistics community detection technique.

Hub	Central Station (1)	Centre (2)	Station Hollands Spoor (3)	Leidschenveen (4)	Leyenburg (5)	Brouwersgracht (6)
# lines	28	21	16	6	10	8
Modularity						
Flow AM	0.28	0.27	0.13	0.04	0.17	0.12
Flow PM	0.29	0.22	0.28	0.04	0.25	0.40
Waiting AM	0.29	0.25	0.13	0.06	0.13	0.16
Waiting PM	0.28	0.21	0.30	0.06	0.26	0.38
# line bundles						
Flow AM	2	2	2	2	2	2
Flow PM	2	2	2	2	3	2
Waiting AM	2	3	2	2	2	2
Waiting PM	2	2	2	2	3	2
% intra-bundle						
Flow AM	0.85	0.77	0.83	0.59	0.75	0.95
Flow PM	0.82	0.77	0.79	0.54	0.59	0.93
Waiting AM	0.86	0.61	0.82	0.58	0.75	0.95
Waiting PM	0.82	0.75	0.80	0.57	0.60	0.94

we performed the line bundle identification for 6 * 4 cases in total: for the AM and PM peak, using transfer flow and transfer waiting as link weight (see Table 5). Table 6 shows the number of unidirectional lines, the modularity value q resulting from this community detection technique (Eq. (10)), the number of line bundles (communities) identified, and for each hub the within-community transfer flow as percentage of the total hub transfer flow. The latter, in Table 6 indicated as '% intra-bundle', is used as additional indicator for the clustering performance.

The modularity value remains at a similar level for the hubs Central Station (1), Centre (2) and Leidschenveen (4), regardless the time period or link weight used. For the hubs Hollands Spoor (3), Leyenburg (5) and Brouwersgracht (6), the modularity is relatively insensitive to the link weight used (transfer flow or transfer waiting time). This can be explained by the relatively high and similar frequency of most urban PT services for the considered hubs. This makes the use of transfer waiting time less distinctive from the use of transfer flow as link weight. However, for these hubs 3, 5 and 6, substantial differences in modularity can be observed between AM and PM, indicating different transfer patterns between these time periods. For the majority of the scenarios two different line bundles are identified. Only for the hub Centre (AM, based on transfer waiting time link weight) and Leyenburg (PM), three line bundles are identified. The percentage within-community transfer flow ranges between 54% and 95% over all scenarios. For most hubs, this percentage is rather stable when a different time period or link weight is used. Similar to the modularity value, this percentage shows to be more sensitive to the time period than the assigned link weight, particularly for the Leyenburg hub 5.

In Fig. 10 the results of the community detection algorithm are visualized for all 24 cases. Each plot shows the lines with their corresponding direction (north-, south-, east- or westbound) which are grouped together as one line bundle. The link width represents the magnitude in terms of transfer flow or transfer waiting time. These case study results display which bundles of lines should be prioritized simultaneously when devising tactical and real-time synchronization measures. Generally, the results show to be intuitive with lines headed in the same direction(s) being grouped together, while producing clusters that could not be formed merely based on grouping each direction. Also in line with expected travel patterns, lines headed in opposite directions (e.g. west- and eastbound lines, or north- and southbound lines) are generally not grouped together. For Central Station (hub 1), one line bundle clearly reflects south-/west-bound passenger journeys, whereas the other bundle reflects north/east-bound journeys. For hubs Centre and Station Hollands Spoor (2 and 3), there is a dominance of northbound and westbound lines being grouped together, and southbound and eastbound lines grouped together. For Leyenburg (hub 5) in the AM, two separate line bundles with eastbound and westbound lines can be detected. During the PM, it can be seen that the westbound lines are grouped into two separate clusters. Probably due to a different mixture of passengers and their corresponding trip purpose and travel patterns, a separate line bundle can be detected between bus lines 23 and 26 in the westbound direction during the PM. At Leidschenveen (hub 4), there is a dominant transfer flow between intersecting tram line 19 southbound and particularly tram line 4 westbound in the AM. However, during the PM a large transfer flow can also be observed between line 19 southbound and line 4 eastbound. At Brouwersgracht (hub 6) is a clear transfer flow between eastbound bus line 25 – from a residential area headed for the city centre – and the tram corridor served by lines 2, 3 and 4 towards the central train station. During the PM an opposite pattern can be observed, with a clear line bundle consisting of tram lines 2, 3 and 4 and westbound bus line 25 bound for a residential area of the city.

5. Conclusions

In this research, we developed a data-driven, generic and passenger-oriented methodology for systematically determining where in the network, and for which lines transfer synchronization should be prioritized in the TSP, so that the TSP becomes solvable for larger, real-world urban PT networks. Our study thus introduces two steps preceding solving the TSP – identify key priorities (a) where to synchronize, and (b) which lines to synchronize. To this end, our method identifies hubs and their spatial boundaries in

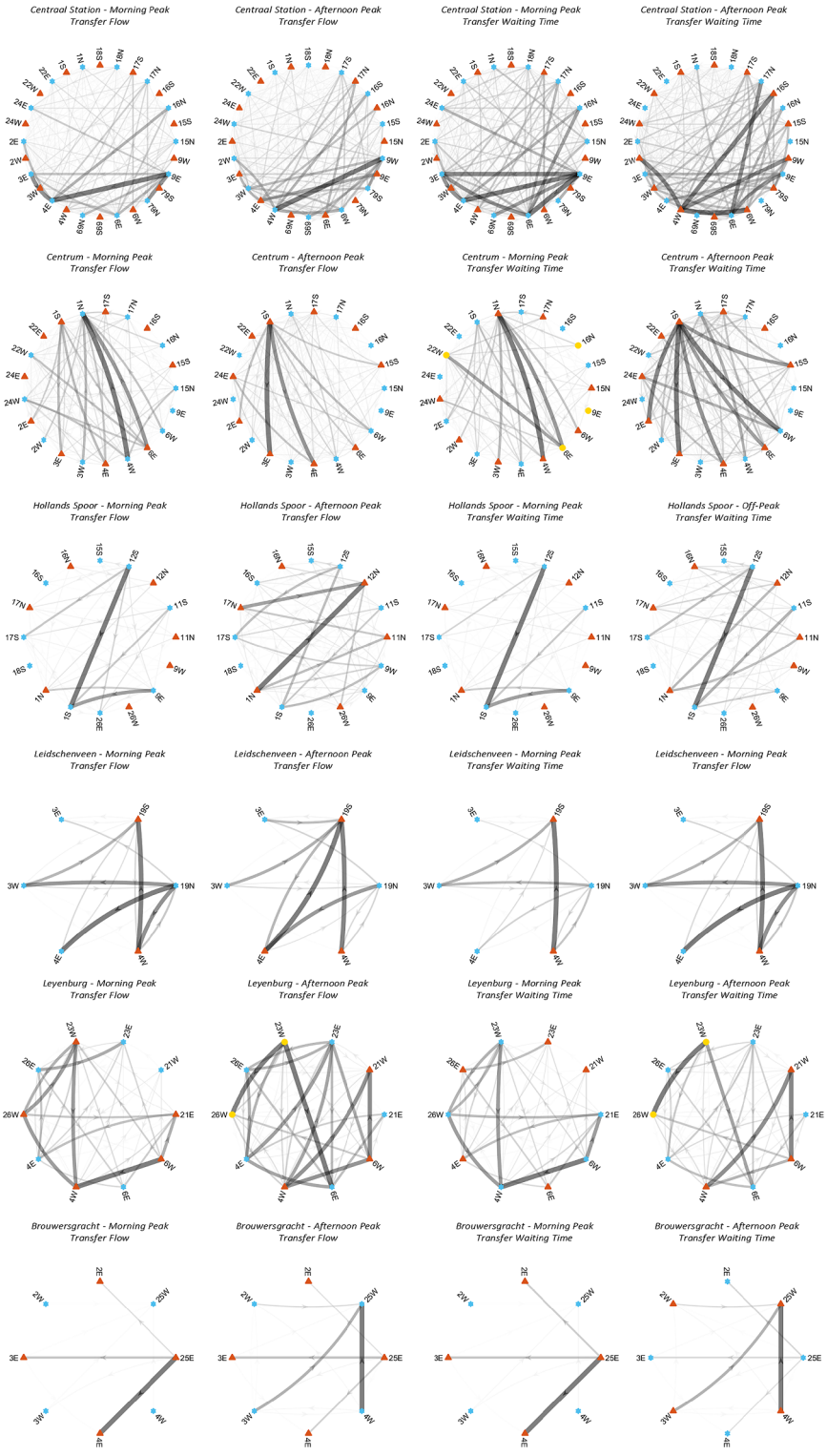


Fig. 10. Results line bundle identification for 6 identified hubs.

urban public transport networks, and determines line bundles within these hubs to be prioritized simultaneously when devising transfer synchronization measures at either tactical or operational planning phases. The proposed non-supervised learning techniques enable the identification of hubs and line bundles based on passenger transfer flows, independent from local knowledge or the geographic location of the urban public transport stops. Our results show that hubs can be composed of a different set of stops when

applying DBSCAN clustering, compared to the set which would result when clustered purely based on geographical information or public stop name. Our clustering results shape the spatial boundaries of public transport transfer locations as used and experienced by passengers. The application of a modularity-based community detection technique shows intuitive lines being grouped together to prioritize during transfer synchronization. Our results illustrate the necessity of synchronizing different line bundles during different periods of the day, depending on the travel patterns prevailing during the relevant time period. The clustering results show to be relatively insensitive to the use of passenger transfer flows or transfer waiting times as link weight, which can be explained by the relatively similar headways associated with all urban PT lines serving a certain hub. If lines with more varying headways would be clustered, a higher sensitivity of clustering results to the used link weight can be expected. Our methodology and study results support public transport operators in timetable design and real-time control, such as holding, by determining where and which lines to synchronize. Moreover, public transport agencies can use these study results to determine where to invest in measures for improving the design of a seamless transfer experience (e.g. amenities, physical environment, island vs. side platforms). Contrary to simply prioritizing pairs of lines with the largest transfer flow between them in the synchronization process, our partitioning approach yields different bundles of lines which should be synchronized simultaneously.

Our approach is able to capture more than 70% of all transfers within identified transfer locations, while only requiring 0.9% of these transfer locations, thus reducing the complexity of solving the TSP substantially at a relatively low cost. In a next step, the optimization process to solve the TSP can be applied to this subset of transfer locations and lines from the total considered real-world PT network. As input to the TSP at the tactical planning phase, our method ranks the transfer locations and the lines at these locations to be prioritized based on passenger flow data. Depending on the network characteristics, the number of lines serving the different transfer locations, the used optimization method and accepted computation time, a PT operator can select the t most important transfer locations with corresponding line bundles to incorporate in the TSP, so that the TSP is solvable within acceptable computation time. The number of transfer locations which can be considered simultaneously may have to be constrained in the TSP in order to make the TSP solvable depending on the approach used. When adopting an approach which relaxes the synchronization to allow for a pre-defined time window (Ibarra-Rojas and Rios-Solis, 2012) or minimizes the total passenger transfer time without pre-setting synchronization requirements (Knoppers and Muller, 1995), the number of transfer locations of interest does not have to be constrained for mid-size urban networks. In case of real-time transfer synchronization decisions in response to an early or late arrival of a PT service of line l_i at a certain transfer location t_i, our method proposes for which lines – clustered within the same line bundle – transfer synchronization should be considered. In a next stage, the optimal holding time decision for different lines can be taken by minimizing the predicted additional travel time for all affected passenger segments, such as transferring passenger, downstream waiting passengers, and downstream transferring passengers (see Gavriilidou and Cats, 2018). In this control framework the predicted impact of synchronization at considered transfer location t_i on potentially missed transfers at downstream transfer locations $t_{j\neq i}$ can be incorporated.

We formulate four recommendations for further research. First, we recommend coupling the optimization process to an assignment model or variable demand model, particularly for networks with relatively many low frequent services. Since passenger demand depends on the quality of the public transport supply, the results from the transfer synchronization following the identified synchronization priorities may influence passenger route choice and, possibly, mode choice. This can result in changes in passenger transfer flows, which in turn can re-set the synchronization priorities. Especially if PT frequencies are relatively low, substantial changes in transfer flows may result from the synchronization process. In particular for PT networks with lower frequencies, it is therefore recommended to couple an assignment model or variable demand model to the optimization process into an iterative supply setting – demand forecasting approach. Second, we recommend experimenting with different clustering techniques in the hub identification phase of our proposed methodology. We used DBSCAN as a density-based, partial clustering technique without a predefined number of clusters, which requires two different input parameters, namely ϵ and θ. Contrary to θ, which can be obtained from the context of the application, ϵ needs to be determined from the θ-distance plot by applying DBSCAN for a large number of instances. Therefore, we recommend testing and comparing the use of other techniques such as OPTICS, in which no distance parameter ϵ needs to be specified explicitly, thus potentially attaining computational gains (Tan et al., 2004). Third, we recommend extending the line bundle identification phase in our study by applying a link-based clustering technique rather than node-based clustering. In our modularity-based community detection technique, the nodes – i.e. lines in a certain direction – are clustered. However, when the transfer links between nodes would be clustered, one would be able to distinguish between transfer flows from line l_i in direction a to l_j in direction b, and flows from l_j in direction b to l_i in direction a. Incorporating the transfer direction between two lines, next to the lines itself, enables deriving further recommendations for timetable planning and real-time coordination by specifying the desired sequence of arrivals. Fourth, further developments may examine how properties of the optimization process, such as choices related to the type of optimization method and type of graph representation, can assist the settings of our methodology.

Acknowledgements

This research was performed as part of the TRANS-FORM (Smart transfers through unravelling urban form and travel flow dynamics) project funded by NWO grant agreement 438.15.404/298 as part of JPI Urban Europe ERA-NET CoFound Smart Cities and Communities initiative. The second author acknowledges the support of the SETA project funded by the European Union's Horizon 2020 research and innovation program. The authors thank HTM, the urban public transport operator of The Hague, the Netherlands, for their valuable cooperation and data provision.

References

Agard, B., Morency, C., Trépanier, M., 2007. Mining public transport user behaviour from smart card data. CIRRELT-2007-42, Canada.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech.: Theory Exp.

Cats, O., Wang, Q., Zhao, Y., 2015. Identification and classification of public transport activity centres in Stockholm using passenger flow data. J. Transp. Geogr. 48, 10–22.

Costa, T.F.G., Lohmann, G., Oliveira, A.V.M., 2010. A model to identify airport hubs and their importance to tourism in Brazil. Res. Transport Econ. 26, 3–11.

Desaulniers, G., Hickman, M.D., 2007. Public transit. In: Barnhart, C., Laporte, G. (Eds.), Handbook in OR & MS. Elsevier, Amsterdam, the Netherlands, pp. 69–127.

Ding, C., Wang, D., Ma, X., Li, H., 2016. Predicting short-term subway ridership and prioritizing ist influential factors using gradient boosting decision trees. Sustainability 8, 1–16.

El Mahrsi, M.K., Come, E., Oukhellou, L., Verleysen, M., 2017. Clustering smart card data for urban mobility analysis. IEEE Trans. Intell. Transp. Syst. 18, 712–728.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. University of Munich, Institute for Computer Science, Germany.

Fortunato, S., Hric, D., 2016. Community detection in networks: a user guide. Phys. Rep. 659, 1–44.

Gavriilidou, A., Cats, O., 2018. Reconciling transfer synchronization and service regularity: real-time control strategies using passenger data. Transportmetrica A (in press).

Gkiotsalitis, K., Cats, O., 2018. Reliable frequency determination: incorporating information on service uncertainty when setting dispatching headways. Transp. Res. Part C 88, 187–207.

Gkiotsalitis, K., Wu, Z., Cats, O., 2019. A cost-minimization model for bus fleet allocation featuring the tactical generation of short-turning and interlining options. Transp. Res. Part C 98, 14–36.

Gordon, J.B., Koutsopoulos, H.N., Wilson, N.H.M., Attanucci, J.P., 2013. Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. Transport. Res. Rec.: J. Transport. Res. Board 2343, 17–24.

Hadas, Y., Ceder, A., 2010. Optimal coordination of public transit vehicles using operational tactics examined by simulation. Transp. Res. Part C 18, 879–895.

Hänseler, F.S., Bierlaire, M., Scarinci, R., 2016. Assessing the usage and level-of-service of pedestrian facilities in train stations: a Swiss case study. Transp. Res. Part A 89, 106–123.

Ibarra-Rojas, O.J., Rios-Solis, Y.A., 2012. Synchronization of bus timetabling. Transp. Res. Part B 46, 599–614.

Ingvardson, J.B., Nielsen, O.A., Raveau, S., Nielsen, B.F., 2018. Passenger arrival and waiting time distributions dependent on train service frequency and station characteristics: a smart card data analysis. Transp. Res. Part C 90, 292–306.

Knoppers, P., Muller, T., 1995. Optimized transfer opportunities in public transport. Transport. Sci. 29, 101–105.

Lancichinetti, A., Fortunato, S., 2009. Community detection algorithms: a comparative analysis. Phys. Rev. E– Stat., Nonlin., Soft Matter Phys. 80, 56117.

Lee, A., van Oort, N., van Nes, R., 2014. Service reliability in a network context: impacts of synchronizing schedules in long headway services. Transport. Res. Rec.: J. Transport. Res. Board 2417, 18–26.

Li, Y., Wang, X., Sun, S., Ma, X., 2017. Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks. Transp. Res. Part C 77, 306–328.

Luo, D., Cats, O., van Lint, J.W.C., 2017. Constructing transit origin-destination matrices with spatial clustering. Transport. Res. Record: J. Transport. Res. Board 2652, 39–49.

Ma, X., Wu, Y., Wang, Y., Chen, F., Liu, J., 2013. Mining smart card data for transit riders' travel patterns. Transp. Res. Part C 36, 1–12.

Munizaga, M., Palma, C., 2012. Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from Santiago, Chile. Transp. Res. Part C 24, 9–18.

Nesheli, M.M., Ceder, A., 2015. A robust, tactic-based, real-time framework for public transport transfer synchronization. Transp. Res. Part C 60, 105–123.

Newman, M.E.J., 2004. Analysis of weighted networks. Phys. Rev. E– Stat. Phys., Plasmas, Fluids, Related Interdiscipl. Top. 70, 9.

Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. Phys. Rev. E– Stat., Nonlin., Soft Matter Phys. 69, 26113.

Nunes, A.A., Dias, T.G., eCunha, J.F., 2016. Passenger journey destination estimation from automated fare collection system data using spatial validation. IEEE Trans. Intell. Transp. Syst. 17, 133–142.

Schakenbos, R., La Paix, L., Nijënstein, S., Geurs, K., 2016. Valuation of a transfer in a multimodal public transport trips. Transp. Pol. 46, 72–81.

Tan, P.N., Steinbach, M., Kumar, V., 2004. Cluster analysis: basic concepts and algorithms. In: Kantardzic, M. (Ed.), Chapter 8, Introduction to Data Mining. Wiley Press.

Trépanier, M., Tranchant, N., Chapleau, R., 2007. Individual trip destination estimation in a transit smart card automated are collection system. J. Intell. Transport. Syst. 11, 1–14.

Van Hagen, M., 2011. Waiting experience at train stations. University of Twente, the Netherlands (PhD Thesis).

Von Ferber, C., Holovatch, T., Holovatch, Y., Palchykov, V., 2009. Public transport networks: empirical analysis and modeling. Eur. Phys. J. B 68, 261–275.

Van Oort, N., Brands, T., de Romph, E., Yap, M.D., 2016. Ridership evaluation and prediction in public transport by processing smart card data: a Dutch approach and example. In: Kurauchi, F., Schmöcker, J.D. (Eds.), Chapter 11, Public Transport Planning With Smart Card Data. CRC Press.

Varga, B., Tettamanti, T., Kulcsár, B., 2018. Optimally combined headway and timetable reliable public transport system. Transp. Res. Part C 92, 1–26.

Wei, Y., Chen, M., 2012. Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks. Transp. Res. Part C 21, 148–162.

Yap, M.D., Cats, O., van Oort, N., Hoogendoorn, S.P., 2017. A robust transfer inference algorithm for public transport journeys during disruptions. Transp. Res. Proc. 27, 1042–1049.

Yildirimoglu, M., Kim, J., 2018. Identification of communities in urban mobility networks using multi-layer graphs of network traffic. Transp. Res. Part C 89, 254–267.

Zhao, J., Rahbee, A., Wilson, N.H.M., 2007. Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. Comput. – Aided Civ. Infrastruct. Eng. 24, 376–387.