

## Estimation of metro network passenger delay from individual trajectories

Krishnakumari, Panchamy; Cats, Oded; van Lint, Hans

**DOI**

[10.1016/j.trc.2020.102704](https://doi.org/10.1016/j.trc.2020.102704)

**Publication date**

2020

**Document Version**

Final published version

**Published in**

Transportation Research Part C: Emerging Technologies

**Citation (APA)**

Krishnakumari, P., Cats, O., & van Lint, H. (2020). Estimation of metro network passenger delay from individual trajectories. *Transportation Research Part C: Emerging Technologies*, 117, Article 102704. <https://doi.org/10.1016/j.trc.2020.102704>

**Important note**

To cite this publication, please use the final published version (if applicable). Please check the document version above.

**Copyright**

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

**Takedown policy**

Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.

Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

# Transportation Research Part C

journal homepage: [www.elsevier.com/locate/trc](http://www.elsevier.com/locate/trc)

## Estimation of metro network passenger delay from individual trajectories

Panchamy Krishnakumari\*, Oded Cats, Hans van Lint

Department of Transport and Planning, Delft University of Technology, Stevinweg 1, 2628CN Delft, the Netherlands



### ARTICLE INFO

#### Keywords:

Smart card data  
Passenger delay  
Matrix solution

### ABSTRACT

Smart card data enables the estimation of passenger delays throughout the public transit network. However, this delay is measured per passenger trajectory and not per network component. The implication is that it is currently not possible to identify the contribution of individual system components – stations and track segments – to overall passenger delay and thus prioritize investments and disruption management measures accordingly. To this end, we propose a novel method for attributing passenger delays to individual transit network elements from individual passenger trajectories. We decompose the delay along a passenger trajectory into its corresponding track segment delay, initial waiting time and transfer delay. Using these delay components, we construct a solvable system of equations, using which the delays on each network component can be computed. The estimation method is demonstrated on one year of data from the Washington DC metro network. Our approach produces promising results by compressing millions of individual trajectories into 3D networks, leading to a dimensionality reduction of 94%. Moreover, the mean slack variable value (that can be interpreted as proxies for estimation errors) is smaller than five seconds per passenger, and has the desired positive sign for almost 90% of all travelers. Applications using the estimation results include revealing network-wide recurrent delay patterns, modeling delay propagation and detecting disruptions.

### 1. Introduction

Service reliability is known to be one of the most important determinants of transit performance, ridership and user satisfaction. Traditionally, service reliability has been focused on vehicles rather than customers using measures such as headway adherence or schedule punctuality (Trépanier et al., 2009). However, there is a strong shift towards using passenger-oriented measures for quantifying the transit performance (Zhao et al., 2013; Hendren et al., 2015). The data with which many of these quantities can be directly or indirectly estimated are already collected by many transport authorities through different sensors and information sources. Some of the well-known and increasingly used data sources for public transport network are infrastructure (stations and track segments) and service network (line) information, timetable data, automatic vehicle location (AVL) data which contains real-time locations of transit vehicles (Moreira-Matias et al., 2015) and by implication the realization of the schedule and automatic fare collection (AFC) or smart card data with origin–destination specific information of passengers (Pelletier et al., 2011).

An increasing body of literature is available on the use of smart card data in public transit; reviews of which can be found in e.g. Pelletier et al. (2011) and Koutsopoulos et al. (2017). Most of these studies focus on origin–destination matrix estimation (Nassir et al., 2011; Gordon et al., 2013), extracting passenger patterns (Ma et al., 2013; Bhaskar and Chung, 2014; Ma et al., 2017), network

\* Corresponding author.

E-mail addresses: [p.k.krishnakumari@tudelft.nl](mailto:p.k.krishnakumari@tudelft.nl) (P. Krishnakumari), [o.cats@tudelft.nl](mailto:o.cats@tudelft.nl) (O. Cats), [j.w.c.vanlint@tudelft.nl](mailto:j.w.c.vanlint@tudelft.nl) (H. van Lint).

<https://doi.org/10.1016/j.trc.2020.102704>

Received 30 September 2019; Received in revised form 21 May 2020; Accepted 10 June 2020

Available online 07 July 2020

0968-090X/© 2020 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

performance analysis (Ma et al., 2015; Ma et al., 2017) and passenger route choice determinants (Yap et al., 2018). AFC systems can be classified based on: (i) whether they include tap-in only or tap-in and tap-out records; (ii) whether fare validation is performed upon boarding (and possibly alighting) the transit vehicle or upon entering (and possibly leaving) a transit station. Depending on the two former aspects, destination (for boarding only) and vehicle (for station validation) inferences might be applied. Methods have been developed to infer the alighting station of a given tap-in record (Trépanier et al., 2007; Sánchez-Martínez, 2017). In addition, new methods have been proposed for the train inference of each passenger, referred as Passenger-to-Train-Assignment (Zhao et al., 2016; Zhu et al., 2017).

By comparing the timetable and AVL data, the vehicle delay of the individual transit vehicles can be directly determined. However, in a shift towards passenger-oriented measures, it is more interesting and relevant to investigate the *passenger delay* for a given line, network segment or transfer station as it encompasses the delay incurred by a passenger within the entire public transit system and not just the on-board delay. Clearly, with the increasing availability of AFC data and progress in Passenger-to-Train-Assignment research, it is possible to estimate and quantify the delay experienced by each passenger. However, it remains unknown how much of the passenger delay can be attributed to individual network elements. In particular, in the context of transit, delays can be associated with different travel time segments - initial waiting time, on-board time and transfer times.

To this end, we study passenger delays on individual network elements by fusing AVL and AFC data. The key contribution of this paper thus is a new data-driven method to derive PT network delays from individual trajectories. The methodology has similarities with the data-driven OD estimation method we propose in Krishnakumari et al., in that we construct a solvable system of equations utilizing all the information at hand without making more assumptions than strictly needed. The usage of Passenger-to-Train-Assignment data makes this study unique because this type of information is relatively new. To the authors' knowledge, this is the first study to explore the potential application of this unique data set. In this work, we distinguish two different types of passenger delay in relation to the public transit network: average passenger delay and total passenger delay. We define the *average passenger delay* as the delay incurred by a passenger while traversing a track segment (link), station (node) or trajectory. The *total passenger delay* is the total delay experienced by all the passengers that traverse that link, node or trajectory. Thus, the total passenger delay is a function of the number of passengers that traverse that network element during a given time period.

In the remainder of this paper, we show that realized passenger-trajectories (resulting from AFC data and Passenger-to-Train-Assignment) and schedule information are sufficient for estimating average and total passenger delays for all the different network elements. The estimation improves by adding a constraint derived from AVL data, because these are—in our case—readily available. With these constraints for each passenger and each vehicle, a solvable system of equations can be formulated. There are various applications for the resulting network indicators, such as identifying key bottlenecks and critical network elements, prioritizing investments and maintenance of assets such as switches, modeling delay propagation through the network and an automated disruption detection. We demonstrate the estimation framework for the metro network of Washington DC using one year of data.

The paper is organized as follows: Section 2 describes the overall estimation framework; in Section 3 we apply the framework on the Washington DC metro network. We outline the network and data used in this section and present the estimation results. We offer conclusions and a discussion on further research avenues in Section 4.

## 2. Methodology

For convenient reference, the notation used for recurrent variables in the methodology is first presented as follows:

### 2.1. Problem formulation

A passenger trajectory or journey  $r_{s_o, s_d, n}$  between origin station  $s_o$  and destination station  $s_d$  of passenger  $n \in \{1, \dots, N\}$ , where  $N$  is the total number of passengers, is defined based on two sets, one of stops and one of lines, each of which are a subset of  $S$  and  $L$ , respectively. The combination of which allows one to define a third set: the set of track segments (subset of  $E$ ) that is traversed by the passenger. As a result, we obtain the initial stop (first element in the stop set), the intermediate transfer stops (second to one before last element in the stop set) and the link set. The stop set is defined as  $\{s_1, \dots, s_i, s_{i+1}, \dots, s_m\}$ , where  $s_2, \dots, s_i, \dots, s_{m-1}$  are the  $m$  transfer stations for that journey, in which  $s_1$  is the origin station  $s_o$  and  $s_m$  is the destination station  $s_d$ . The line set for the same journey is defined as  $\{l_1, \dots, l_i, \dots, l_{m-1}\}$ .

In the following, we assume that the origin and destination station, and the trajectory  $r_{s_o, s_d, n}$  of each passenger are known. The passenger trajectory contains all the trip-legs for a given passenger journey with the corresponding time stamps such as the tap-in, tap-out and transfer information. Depending on the fare validation scheme, these variables can either be directly observed or need to be inferred. If the destination of the passenger is not known, we can apply the by now well-established destination inference methods (Trépanier et al., 2007; Sánchez-Martínez, 2017). If the trajectory of the passenger is not known, one can employ trajectory inference methods such as Passenger-to-Train-Assignment (Zhu et al., 2017) and ODX method (Sánchez-Martínez, 2017). In case of vehicle-based tap-validation, there is need to also determine whether two subsequent transactions constitute parts of a single journey by performing a transfer inference (Yap et al., 2017). Depending on the fare validation scheme, inferences (destination, trajectory and/or transfer) are performed so that the trajectory  $r_{s_o, s_d, n}$  of each passenger is given as input to the method proposed in this study, so that the passenger delay  $d_{s_o, s_d, k, n}$  is defined as:

$$d_{s_0,s_d,k,n} = \begin{cases} t_{s_0,s_d,k,n} - \tilde{t}_{s_0,s_d,k} & \text{if } t_{s_0,s_d,k,n} > \tilde{t}_{s_0,s_d,k} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where  $t_{s_0,s_d,k,n}$  is the observed travel time for passenger  $n$  departing at time period  $k$  and can be obtained by finding the difference between the tap-in and tap-out time of that passenger. Any negative delay  $d_{s_0,s_d,k,n}$  is set to 0 as shown in (1), since this implies the passenger reached the destination earlier than expected.

The maximum scheduled travel time  $\tilde{t}_{s_0,s_d,k}$  for a journey  $r_{s_0,s_d}$  can be defined in several ways. In this study, the scheduled travel time include walking times (at both ends as well as at the transfer location if any), maximum waiting times assumed to equal the service headway (of the first leg and any other leg in case of transfers) and the scheduled on-board travel time between boarding and alighting stations (for each leg). These values were made available by the local public transport authority. We then formally define the scheduled travel time as follows:

$$\tilde{t}_{s_0,s_d,k} = \sum_{l \in r_{s_0,s_d}} \sum_{s_i \in l} \tilde{t}_{s_i,s_{i+1},k}^{veh} + \sum_{l \in r_{s_0,s_d}} h_{l,k} + \sum_{s_i \in r_{s_0,s_d}} \sum_{s_j=s_1}^{s_m} t_{s_i}^{walk} \quad (2)$$

where  $\tilde{t}_{s_0,s_d,k}$  is composed of scheduled running times and dwell times  $\tilde{t}_{s_i,s_{i+1},k}^{veh}$  assigned to track segments connecting subsequent stations  $s_i$  and  $s_{i+1}$ , headway  $h_{l,k}$  between successive services for the lines  $l$  in that journey and walking time  $t_{s_i}^{walk}$  at the origin station and transfer stations denoted by  $s_i$ . The headway between transit services is included in calculating the schedule travel time due to the definition of on-time journey considered in this study. Note that a passenger is considered to be on-time with regards to the maximum scheduled journey time even he/she just missed a transit vehicle but was able to catch the next service of that line and the headway between the services account for that fallback time.

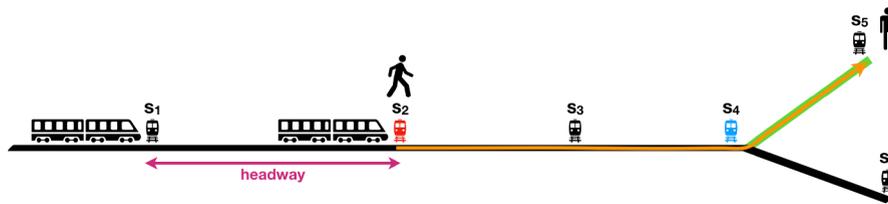
Given the individual trajectories of the passengers, the aim is to decompose the delay experienced on a given passenger journey into the corresponding network elements they traversed along their trajectory. The estimated delays for each network element are assumed common (or average) for all individuals. We assume that the observed travel time of a passenger  $t_{s_0,s_d,k,n}$  comprises of travel time that can be attributed to one of the following network elements of the transit service:

1. Time spent at the origin stop of the journey
2. On-board a transit vehicle along one of the segments
3. Time spent at a transfer station

Similarly, the delay experienced by a passenger is comprised of the delays occurring at these network elements. The composition of the scheduled and observed passenger travel time is illustrated in Fig. 1. Based on this schematic representation, the passenger delay between stations ( $s_0, s_d$ ) for a given departure time  $k$  can be defined as:

$$d_{s_0,s_d,k,n} \geq d_{s_0,k}^{wait} + \sum_{l \in r_{s_0,s_d,n}} \sum_{s_i \in l} d_{s_i,s_{i+1},k+t_{s_0,s_i,k}}^{on-board} + \sum_{s_i \in r_{s_0,s_d,n}} \sum_{s_j=s_1}^{s_{m-1}} d_{s_i,k+t_{s_0,s_i,k}}^{trans} \quad (3)$$

where  $d_{s_0,k}^{wait}$  is the initial waiting delay at the origin station  $s_0$ ,  $d_{s_i,s_{i+1},k+t_{s_0,s_i,k}}^{on-board}$  is the on-board delay between track segment ( $s_i, s_{i+1}$ ) and  $d_{s_i,k+t_{s_0,s_i,k}}^{trans}$  is the transfer delay at transfer stop  $s_i$  in journey  $r_{s_0,s_d,n}$ . The on-board delay and the transfer delay might not be associated with the same time period as the departure time period  $k$ . The corresponding time period for these component can be calculated by adding the departure time period  $k$  to the observed travel time  $t_{s_0,s_i,k}$  between the origin station  $s_0$  and the corresponding origin  $s_i$  of track segment ( $s_i, s_{i+1}$  or transfer station  $s_i$ ). Alternatively  $k + t_{s_0,s_i,k}$  relates to the arrival time at the track segment or the transfer station  $s_i$ . However, we can simplify the delay definition but this might lead to estimation error depending on the temporal aggregation of the time period  $k$ . The simplified delay definition is as follows:



**scheduled travel time = on-board time + headway + walking time**

**observed passenger travel time = waiting time + on-board time + transfer time**

**passenger delay = observed passenger travel time - scheduled travel time**

Fig. 1. Schematic picture of scheduled and observed travel time between station  $s_2$  and  $s_5$ .

$$d_{s_0, s_d, k, n} \geq d_{s_0, k}^{wait} + \sum_{l \in r_{s_0, s_d, n}} \sum_{s_i \in l} d_{s_i, s_{i+1}, k}^{on-board} + \sum_{s_i \in r_{s_0, s_d, n}} \sum_{s_i = s_1}^{s_m-1} d_{s_i, k}^{trans} \quad (4)$$

Based on the type of the fare validation scheme, the passenger delay definition may differ. For example, the initial waiting time delay for a surface bus system cannot be observed from the AFC data and hence, the waiting delay component will then be removed from (4). The delay definition in (4) can be further simplified or expanded based on the specifics of the transit system subject to investigation.

The inequality in (4) is due to potential *non-observable* personal travel components which may add additional delays such as performing an activity at one of the stations within the gated area. While on-board, waiting and transfer delays relate to those delays that can be attributed to service elements, any other delay - including possibly walking time related - is assumed unsystematic and is therefore attributed to the individual. Formulating the relationship between these personal delays and the three network related delay components by means of an inequality constraint allows us to perform the delay inference without ignoring unobserved personal delay components. We define this unobserved error term as a slack variable for each passenger; thus (4) can be reformulated as:

$$d_{s_0, s_d, k, n} = d_{s_0, k}^{wait} + \sum_{l \in r_{s_0, s_d, n}} \sum_{s_i \in l} d_{s_i, s_{i+1}, k}^{on-board} + \sum_{s_i \in r_{s_0, s_d, n}} \sum_{s_i = s_1}^{s_m-1} d_{s_i, k}^{trans} + \epsilon_n \quad (5)$$

An additional constraint can also be formulated, that pertains to the on-board delay component, which must be equal to the transit vehicle delay between the corresponding trip segments. This vehicle delay can be directly inferred from the AVL data and the schedule information as follows:

$$d_{s_0, s_d, k, v}^{veh} = \sum_{l \in r_{s_0, s_d, v}} \sum_{s_i \in l} \begin{cases} t_{s_i, s_{i+1}, k, v}^{veh} - \tilde{t}_{s_i, s_{i+1}, k}^{veh} & \text{if } t_{s_i, s_{i+1}, k, v}^{veh} > \tilde{t}_{s_i, s_{i+1}, k}^{veh} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

This results in an additional set of equations to solve for the unknowns in (5) which reads

$$d_{s_0, s_d, k, v}^{veh} = \sum_{l \in r_{s_0, s_d, v}} \sum_{s_i \in l} d_{s_i, s_{i+1}, k}^{on-board} \quad (7)$$

where  $d_{s_0, s_d, k, v}^{veh}$  is the delay for transit vehicle  $v$  servicing the line that starts at  $s_0$  and ends at  $s_d$  within time  $k$  and  $r_{s_0, s_d, v}$  is the trip composed of track segments visited by transit vehicle  $v$ . The initial waiting time and transfer delay component are irrelevant for a transit vehicle trajectory. The slack variable can also be set to 0 as there is no unobserved delay component for the vehicle delay.

## 2.2. Formulating a solvable system of equations

We decompose each of the passenger delay into delay at three network elements - origin stations, track segments and transfer stations and each transit vehicle delay into delay per track segments. With (5), now each passenger trajectory can be written as a linear combination of these three passenger delay component types with an additional slack variable for the error. Our main hypothesis is that formulating (5) for all passenger trajectories leads to a potentially solvable system of equations, since each passenger's trip between  $k$  and  $k + t_{s_0, s_d, k, n}$  serves as a constraint for all other trips that traverse one or more common network elements during this trip.

We demonstrate this point with an example. Consider the toy network shown in Fig. 1. If a passenger traveling from  $s_1$  to  $s_5$  experiences no delay while a passenger traveling from  $s_1$  to  $s_6$  experiences a delay of say 5 min, then this delay is probably due to delay occurring between stations  $s_4$  and  $s_6$ . Thus, given a sufficient number of passengers, there exists a bounded solution for (5).

Given that  $d_{s_0, s_d, k, n}$  and  $r_{s_0, s_d, n}$  are known, we can expand each delay in (5) to a linear combination of single network elements. Thus, the initial waiting time can be expanded as a linear combination of station elements as follows:

$$d_{s_0, k}^{wait} = \sum_{s_i \in S} \alpha_{s_i, k, n} d_{s_i, k}^{wait} \quad (8)$$

where

$$\alpha_{s_i, k, n} = \begin{cases} 1 & s_i = s_0 \ \forall s_0 \in r_{s_0, s_d, n} \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

The on-board time can be reformulated based on track segment elements:

$$\sum_{l \in r_{s_0, s_d, n}} \sum_{s_i \in l} d_{s_i, s_{i+1}, k}^{on-board} = \sum_{(s_i, s_{i+1}) \in E} \beta_{s_i, s_{i+1}, k, n} d_{s_i, s_{i+1}, k}^{on-board} \quad (10)$$

where

$$\beta_{s_i, s_{i+1}, k, n} = \begin{cases} 1 & (s_i, s_{i+1}) \in r_{s_0, s_d, n} \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

The transfer time can be defined based on the transfer stations:

$$\sum_{s_l \in r_{s_0, s_d, n}} \sum_{s_i = s_l}^{s_m-1} d_{s_i, k}^{trans} = \sum_{s_l \in S^{trans}} \gamma_{s_i, k, n} d_{s_i, k}^{trans} \quad (12)$$

where

$$\gamma_{s_i, k, n} = \begin{cases} 1 & s_i \in r_{s_0, s_d, n} \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

Note that the stations and transfer stations are directed for public transport network. The number of station directions depends on the number of its outgoing neighbors. Thus, for station directed  $s_4$  in Fig. 1, there are three directions possible, either towards  $s_3$ ,  $s_5$  or  $s_6$ . This implies that the number of directed stations in a public transport network is the same as the number of directed edges. We thus refine (8) and (12) based on these directed travel nodes as follows:

$$d_{s_0, k}^{wait} = \sum_{(s_i, s_{i+1}) \in E} \alpha_{s_i, s_{i+1}, k, n} d_{s_i, s_{i+1}, k}^{wait} \quad (14)$$

where

$$\alpha_{s_i, s_{i+1}, k, n} = \begin{cases} 1 & s_i = s_1 = s_0 \quad \forall (s_1, s_2) \in r_{s_0, s_d, n} \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

$$\sum_{s_l \in r_{s_0, s_d, n}} \sum_{s_i = s_l}^{s_m-1} d_{s_i, k}^{trans} = \sum_{(s_i, s_{i+1}) \in E} \gamma_{s_i, s_{i+1}, k, n} d_{s_i, s_{i+1}, k}^{trans} \quad (16)$$

where

$$\gamma_{s_i, s_{i+1}, k, n} = \begin{cases} 1 & (s_i, s_{i+1}) \in r_{s_0, s_d, n} \quad \forall s_i \in S^{trans} \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

Next, we can reformulate the passenger delay given in (5) as:

$$d_{s_0, s_d, k, n} = \sum_{(s_i, s_{i+1}) \in E} \alpha_{s_i, s_{i+1}, k, n} d_{s_i, s_{i+1}, k}^{wait} + \sum_{(s_i, s_{i+1}) \in E} \beta_{s_i, s_{i+1}, k, n} d_{s_i, s_{i+1}, k}^{on-board} + \sum_{(s_i, s_{i+1}) \in E} \gamma_{s_i, s_{i+1}, k, n} \delta_{s_i, s_{i+1}, k}^{trans} + \epsilon_n \quad (18)$$

Similarly, we can formulate the transit vehicle delay given in (7) as:

$$d_{s_0, s_d, k, v}^{veh} = \sum_{(s_i, s_{i+1}) \in E} \beta_{s_i, s_{i+1}, k, v} d_{s_i, s_{i+1}, k}^{on-board} \quad (19)$$

where

$$\beta_{s_i, s_{i+1}, k, v} = \begin{cases} 1 & (s_i, s_{i+1}) \in r_{s_0, s_d, v} \\ 0 & \text{otherwise} \end{cases} \quad (20)$$

Based on the example network from Fig. 1, for a given passenger 1 with his or her journey defined by link set  $\{(s_2, s_3), (s_3, s_4), (s_4, s_5)\}$  and stop set  $\{s_2, s_4, s_5\}$  where  $s_4$  is a transfer station, we can write the expanded form of (18) for passenger 1 as:

$$d_{s_0, s_d, k, 1} = d_{s_2, s_3, k}^{wait} + d_{s_2, s_3, k}^{on-board} + d_{s_3, s_4, k}^{on-board} + d_{s_4, s_5, k}^{on-board} + d_{s_4, s_5, k}^{trans} + \epsilon_1 \quad (21)$$

Similarly, for a given transit vehicle 1 with its journey defined using link set  $\{(s_1, s_2), (s_2, s_3), (s_3, s_4), (s_4, s_5)\}$ , we can write the expanded form of (19) for vehicle 1 as:

$$d_{s_0, s_d, k, 1}^{veh} = d_{s_1, s_2, k}^{on-board} + d_{s_2, s_3, k}^{on-board} + d_{s_3, s_4, k}^{on-board} + d_{s_4, s_5, k}^{on-board} \quad (22)$$

We can generalize (18) and (19) to  $N$  passengers and  $V$  vehicles respectively to build a system of equations and thus a matrix equation as:

$$Cx + \epsilon = B, \quad x \geq 0; \quad (23)$$

where

$$C = \begin{bmatrix} 0 & 1 & \dots & 0 & 1 & \dots & 0 & 0 & \dots \\ \vdots & \ddots & & & & & & & \\ \alpha_{s_1,s_2,k,n} & \alpha_{s_1,s_2,k,n} & \dots & \beta_{s_1,s_2,k,n} & \beta_{s_2,s_3,k,n} & \dots & \gamma_{s_1,s_2,k,n} & \gamma_{s_2,s_3,k,n} & \dots \\ \vdots & \ddots & & & & & & & \\ 0 & 0 & \dots & 1 & 1 & \dots & 0 & 0 & \dots \\ \vdots & \ddots & & & & & & & \\ 0 & 0 & \dots & \beta_{s_1,s_2,k,v} & \beta_{s_2,s_3,k,v} & \dots & 0 & 0 & \dots \\ \vdots & \ddots & & & & & & & \end{bmatrix}; \quad (24)$$

$$\mathbf{x} = \begin{bmatrix} d_{s_1,s_2,k}^{wait} \\ d_{s_2,s_3,k}^{wait} \\ \vdots \\ d_{s_1,s_2,k}^{on-board} \\ d_{s_2,s_3,k}^{on-board} \\ \vdots \\ d_{s_1,s_2,k}^{trans} \\ d_{s_2,s_3,k}^{trans} \\ \vdots \end{bmatrix}; \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \\ \vdots \\ 0 \\ \vdots \\ 0 \\ \vdots \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} d_{s_0,s_d,k,1} \\ \vdots \\ d_{s_0,s_d,k,n} \\ \vdots \\ d_{s_0,s_d,k,1}^{veh} \\ \vdots \\ d_{s_0,s_d,k,v}^{veh} \\ \vdots \end{bmatrix} \quad (25)$$

The  $\mathbf{x}$  is shorthand for the delay attributed to each network element;  $C$  contains ones or zeros (LHS of Eqs. (18) and (19));  $\epsilon$  is the slack variable for each passenger and  $B$  corresponds to the passenger and vehicle delay (RHS of Eqs. (18) and (19)), respectively. A zero in the  $C$  matrix implies that the corresponding network element does not contribute to the respective passenger or vehicle delay in  $B$  and a value of one imply that the corresponding network element is part of that passenger's or vehicle's journey and hence contributes to the delay. Thus, Eqs. (18) and (19) now constitute a solvable system of equations in which  $d_{s_i,s_{i+1},k}^{wait}$ ,  $d_{s_i,s_{i+1},k}^{on-board}$  and  $d_{s_i,s_{i+1},k}^{trans}$  are the unknowns  $\forall (s_i, s_{i+1}) \in E$  and  $d_{s_0,s_d,k,n}$  and  $d_{s_0,s_d,k,v}^{veh}$  are known  $\forall n \in \{1, \dots, N\}$  and  $v \in \{1, \dots, V\}$  respectively. The matrix equality given in Eq. (23) can be solved using a constrained linear least squares solution (Altman and Gondzio, 1999) with the lower bound set to 0 to ensure a non-negative solution as follows:

$$\min_{\mathbf{x}} \frac{1}{2} \left\| C\mathbf{x} - B \right\|_2^2 \quad (26)$$

In case a non-negative solution does not exist, an ordinary least square solution is computed and the negative values of the delay matrix are ignored when computing the estimation error. The minimum requirement to solve the system of equations for a given time period is that at least one passenger traverses each of the network elements.

### 2.3. Evaluation metrics

An optimal solution is reached when  $\epsilon$  is minimized. Thus, we use the slack variable in Eq. (27) to evaluate the estimation results. The delay estimates  $\mathbf{x}$  of the individual network elements are used to reconstruct the individual passenger delay by multiplying  $C$  and  $\mathbf{x}$ . This can be used to find the estimation error or the slack variable as follows:

$$\epsilon = B - C\mathbf{x} \quad (27)$$

This estimation error evaluates the discrepancy between the reconstructed individual total passenger delay and the ground-truth passenger delay at the journey level. Based on our problem formulation, a positive slack variable is favored as this implies that our estimates are less than or equal to the observed delay of the passenger. This is reasonable as some of the delay may be attributed to individual-related factors rather than a delay associated with the respective network element. Conversely, a negative slack variable value implies that our estimation attributes to network elements a total delay that exceeds the delay experienced by a passenger. Consequently, the expected slack distribution is skewed towards positive values whereas the error distribution is assumed to be Gaussian in the least square approach (Hayashi, 2000). However, we choose to employ the least square approach as a first estimation and evaluate if it can reasonably solve our system of equations for our formulation.

## 3. Application

In this section, we demonstrate our estimation method on a real-world application. For this, we first explain the data and network used (Section 3.1). Thereafter, we provide some useful descriptive statistics of the data (Section 3.2) and finally, we present the results of the estimation (Section 3.3) and analyze its performance (Section 3.4).

### 3.1. Data

We demonstrate the estimation approach on a data set of smart card data from the Washington DC Area Metro Transit Authority

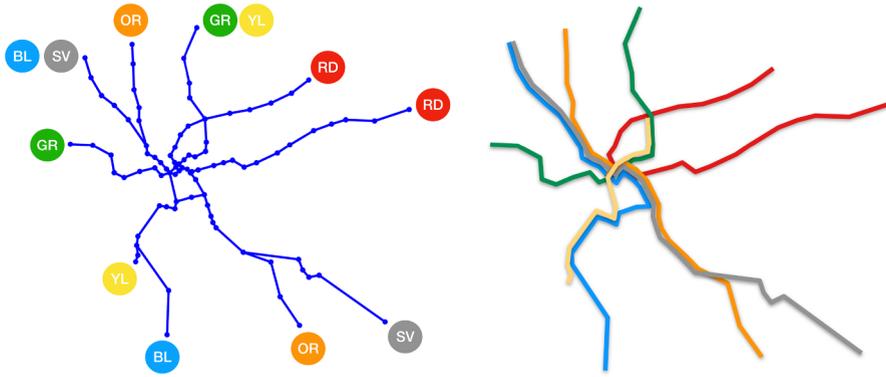


Fig. 2. Washington metro network.

(WMATA) in the United States. The data is composed of one year of smart card data from 19 August 2017 to 28 August 2018 for the entire metro network of Washington DC which contains the Passenger-to-Train-Assignment outputs derived from an application of the so-called ODX method described in Sánchez-Martínez (2017). In general, the ODX method infers the origin, destination and transfer of a given passenger. For Washington metro, the tap-in and tap-out location and time stamp of a passenger are directly available. Therefore, the method only infers the paths and transfers of the (possibly multi-leg) journeys by minimizing a generalized disutility function. The method searches for the path that minimizes the time between the tap-out and the latest possible train arrival time. This is under the assumption that passengers do not engage in non-trip related activities at the destination. Thus, for each passenger, the method provides the following trip legs which includes both the location and the corresponding timestamps access (directly observed), egress (directly observed), ungated transfer and train ride itself. In addition, the dataset also includes the rail movement data, schedule information and disruption log file. The metro network is comprised of 6 lines, 91 stations, 186 links and 9 transfer stations as shown in Fig. 2.

### 3.2. Descriptive statistics

The network has an average ridership of  $\approx 438\,000$  rides per day and a total of  $\approx 157$  million rides during the entire study period with an average journey time of 28 min per passenger. Of these, 14% of the passengers experience a delay. For those experiencing a delay, the mean delay is 6 min. Moreover, 39% of passenger trips include transfers with an average of 1.14 trip legs per passenger journey. A breakdown of the number of passengers for different time periods is shown in Fig. 3. There are two distinct peaks in the passenger distribution - morning and afternoon peak.

### 3.3. Estimation results

In this section, we present the results of the delay estimation for the Washington metro network. We chose a temporal aggregation of 30 min for  $k$  in (18), since the maximum headway between transit vehicles is 20 min and choosing a temporal aggregation smaller than that would imply that there would be no vehicles between some OD pairs, hence no passengers and consequently no system of

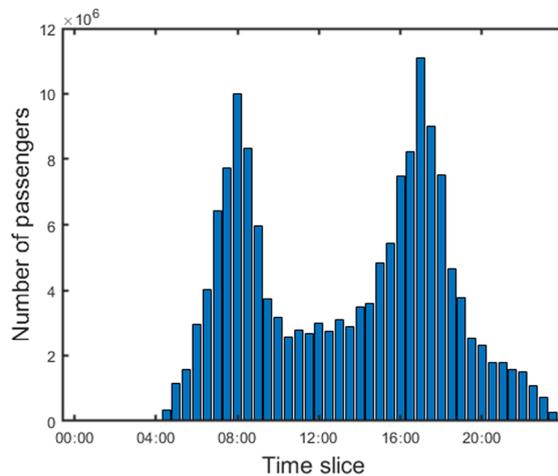
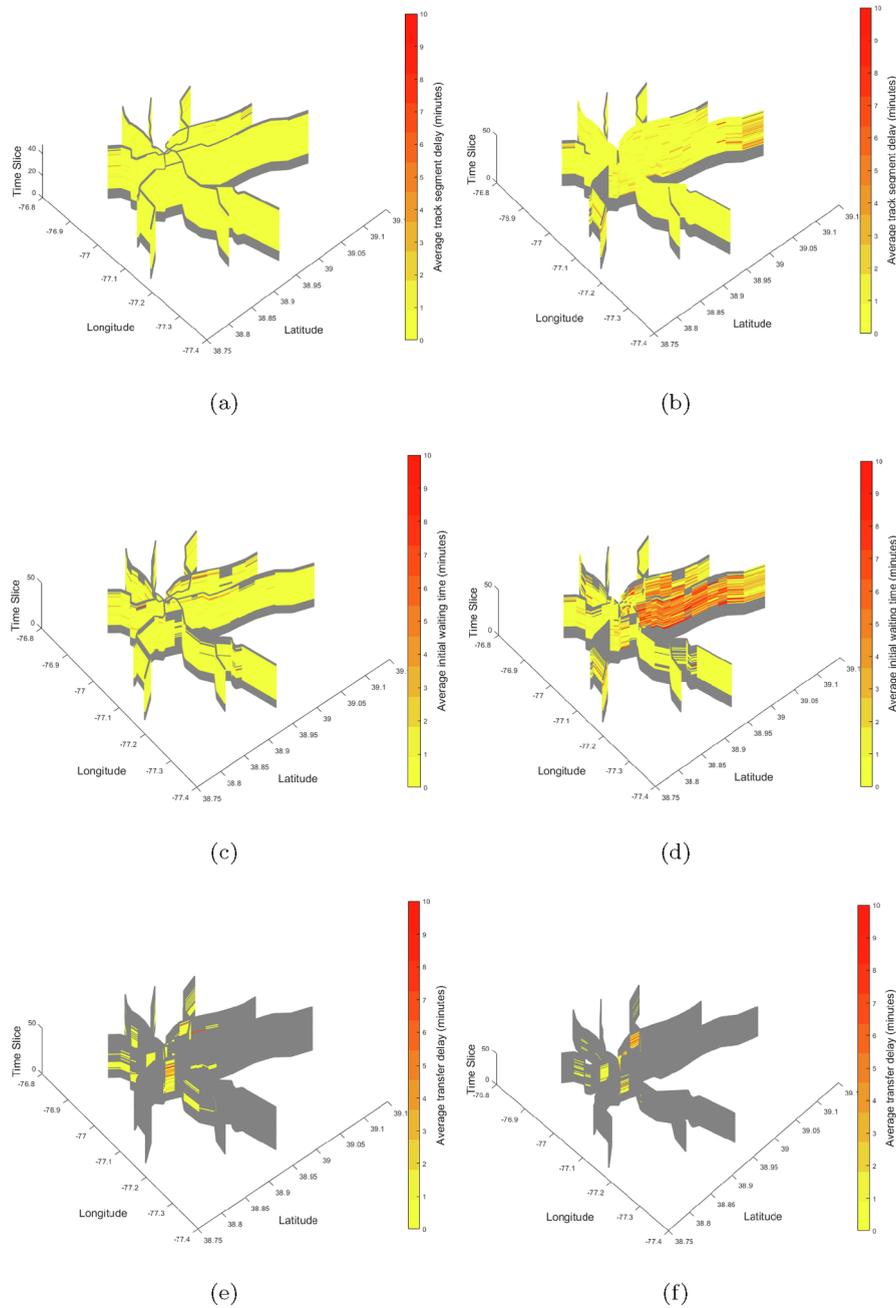


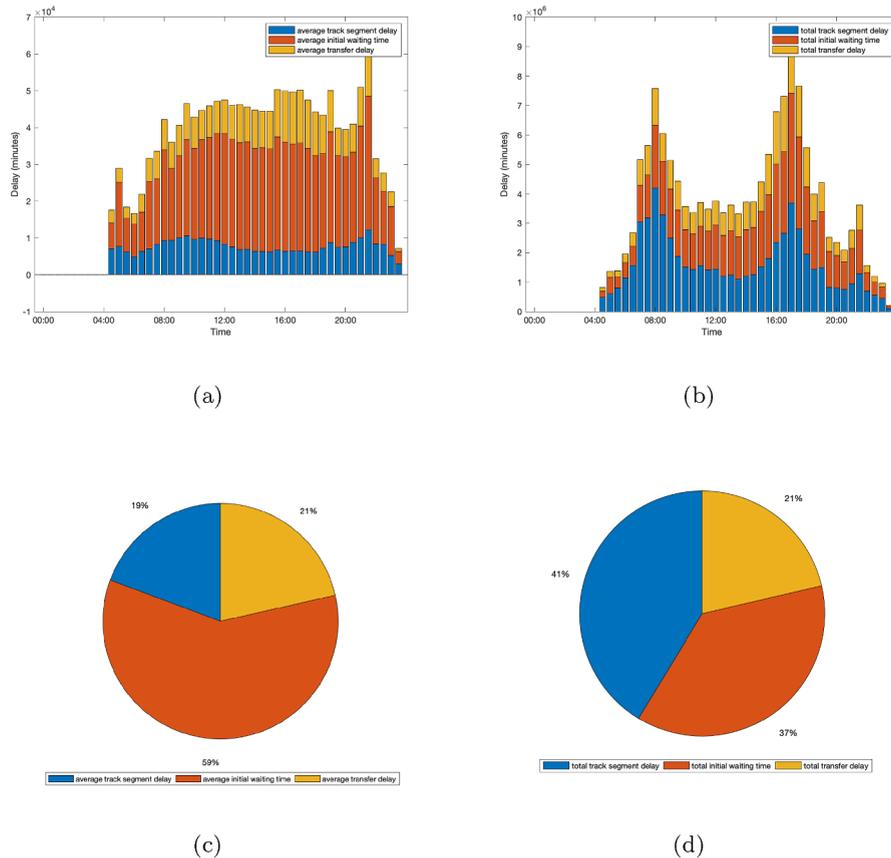
Fig. 3. Passenger distribution of all days for different time periods.



**Fig. 4.** Estimation results of average track segment delay, average initial waiting time and average transfer delay for a weekday dated 01–03-2018 is shown in (a), (c) and (e) respectively and for a weekend day dated 03–03-2018 is shown in (b), (d) and (f) respectively.

equations. Having an aggregation of 30 min ensures that at least one transit vehicle per time period is included in the system of equations as represented in (18).

Fig. 4 shows the estimation results of the average passenger delay for the three network elements for a selected weekday (Thursday), i.e. March 1, 2018, with a temporal aggregation of 30 min. There is no significant track segment delay for this particular day. The waiting time and transfer delay are mapped on a link rather than a node as we incorporate directionality of the node to distinguish journeys between different lines. This allows us to build these three compact 3D graphs for visualizing the delay propagation of each day for each of the delay components. This can be used for evaluating the performance of the metro network on a given day or over a long period of time or estimate the passenger delay incurred between any given origin–destination (OD) pair. Moreover, there were  $\approx 596\ 000$  rides recorded on this particular weekday and we were able to represent the dynamics of the network using these three 3D networks with the dimensions  $3 \times 48 \times 186$  ( $3 \times$  temporal aggregation  $\times$  number of links), thus leading



**Fig. 5.** Delay distribution of the whole dataset across the three network elements for different time periods based on (a) average passenger delay and (b) total passenger delay. Overall delay distribution of the whole dataset across the three network elements based on (c) average passenger delay and (d) total passenger delay.

to a dimensionality reduction of about 95% (26784/596000).

A 3D graph of a weekend day (Saturday), i.e. March 3, 2018, in the same month is also shown in Fig. 4. This exhibits a significant delay in the red line compared to a normal weekday. From the disruption log file of WMATA, we are able to attribute this delay to late track clearing in the morning which had a cascading effect on the rest of the line. One of the main findings from our delay estimation is that not all the delays have a corresponding explanation in the disruption log file of WMATA, which is maintained manually, and vice versa. Thus, our estimation can be used to enrich the log file with additional incidents that caused passenger delay as well as quantify the consequences of these incidents in terms of the number of passengers affected, the average passenger delay and the spatial extent of the impact of the incident.

We applied the estimation framework on the 359 days of smart card data with around 157 million rides and compressed each day into the three 3D delay networks, leading to an overall dimensionality reduction of 94%. We also used the estimates to explore the distribution of the average and total passenger delays for the entire analysis period decomposed into different network elements and different time periods as shown in Fig. 5. The average passenger delay, shown in Fig. 5(a), is stable compared to the morning and evening peak in the total passenger delay distribution in Fig. 5(b). Thus, delays occur at all time periods, whereas more passengers are affected during the peak periods as can be expected. Figs. 5(c) and (d) show the contribution of each network element delay to the overall delay. In the case of average passenger delay, 59% of the delay is associated with the initial waiting time. However, when the number of passengers affected is considered, track segment delay contributes the most with 41% of the delay.

### 3.4. Validation

We evaluate the validity of the estimation by reconstructing the individual passenger delay and calculating the slack variable. A slack value of 0 implies that the delay experienced by a passenger based on the AFC data is the same as obtained by summing over our estimates for the respective journey travel components. We do expect a slight variation in the slack value of each passenger due to heterogeneity in user behavior. This is confirmed by inspecting the distribution of the slack variable using boxplot for the above mentioned selected weekday and weekend day shown in Figs. 6(a) and (b).

In Figs. 6 compact boxplot with small variation is desired as this implies that the delay estimated for all passengers at the time

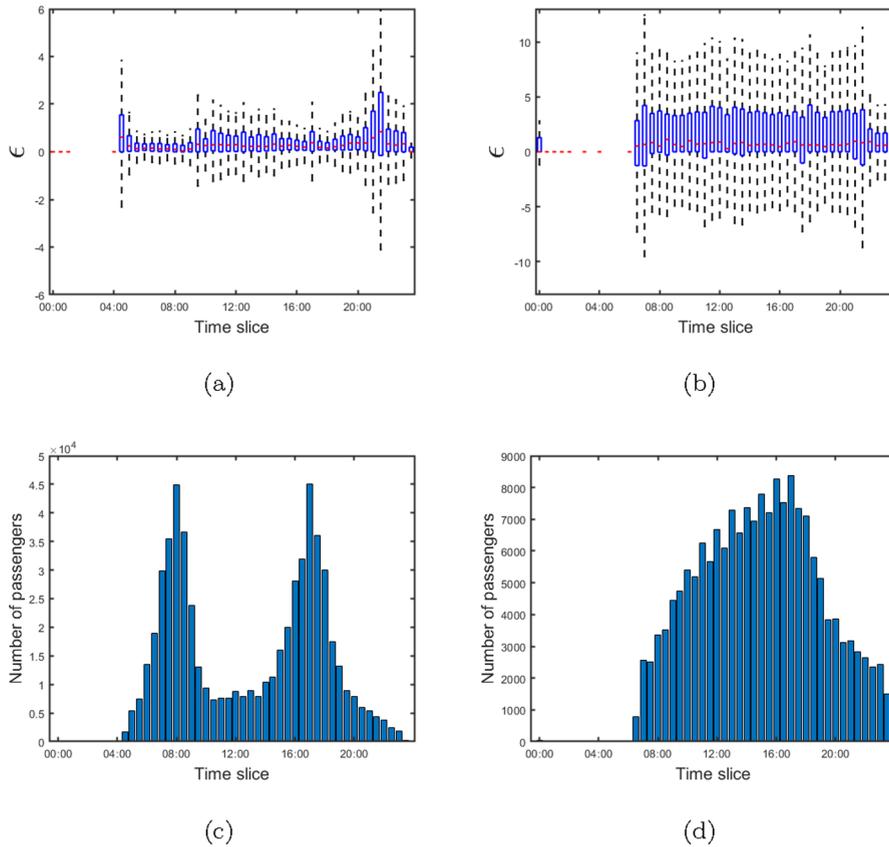


Fig. 6. (a) and (b) Slack value distribution per time period which also shows the dispersion within the time period for a weekday and a weekend respectively; (c) and (d) Passenger distribution per time period for a weekday (01–03-2018) and a weekend day (03–03-2018), respectively.

period falls within that small range of values. Furthermore, the dispersion of the slack variable increases with a decrease in the number of passengers for a given time period which can be seen in Fig. 6 as slack distributions in Figs. 6(a) and (b) are complemented with passenger ridership distributions in Figs. 6(c) and (d). This is expected as more passengers mean more equations and thus more confidence in the estimates and thus a lesser dispersion. The effect of the number of passengers is also evident from the distribution of the slack for the weekday. The weekday has a smaller number of passengers relative to the weekend and thus the boxplot is also relatively compact. Moreover, the distinctively different passenger distribution than the weekday connect to the slack variable result.

Another important finding from Fig. 6(a) and (b) is that some passengers have a negative slack. This is presumably primarily due to the time aggregation. Our estimation is aggregated for a certain  $k$  time interval. Assume that a disruption occurred at  $k + i$  time whereas a passenger departed at time  $< k + i$ , thus avoiding the disruption. However, due to the level of aggregation, that trip will be

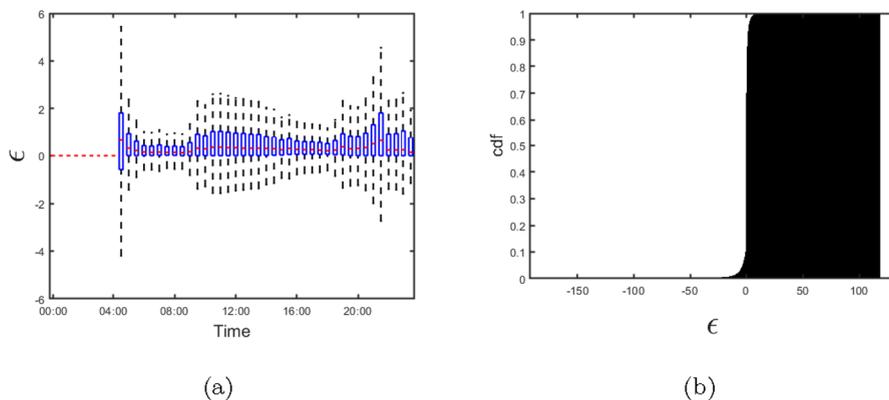


Fig. 7. (a) Overall slack distribution of all passengers across all days for different time period (b) Cumulative distribution function of the slack values for all passengers.

categorized as being affected by the disruption, since most of the passengers in that time period experienced that delay.

The slack distribution for all the days across different time periods is shown in Fig. 7. As can be seen from the figure, most of the values are distributed around zero and our estimation framework achieves a mean slack of 0.072 min across all the passengers which is promising. Additionally, only 10.76% of all passengers had a negative slack value, with a mean of -5 min. Moreover, when complemented with the passenger distribution shown in Fig. 3, it is evident that the boxplot at a time period with a large number of passengers are more compact with less dispersion compared to the time period with less number of passengers. This is, also, in line with the finding from the individual weekday and weekend.

In order to assess the validity of using the least square method to solve the system of equations, we conducted model diagnostics for the error terms, i.e. the slack variables. There are three key assumptions about the error term that are necessary for the least square approach to provide unbiased, efficient and linear estimators: a) errors follow identical distributions, b) errors are independent, c) errors are normally distributed. We obtain a p-value greater than 0.05 for the Breusch-Pagan and Durbin-Watson tests, thus confirming the null hypotheses of homoscedasticity and independence, respectively. Moreover, the Shapiro-Wilk normality test value confirms what we observed in the slack variable distribution that the errors are not normally distributed. It is unclear whether non-normality would substantially affect our results although there is evidence to suggest that when the normality assumption is not valid but the other assumptions are, the estimates are still consistent (Box, 1976). Different approaches can be taken to remedy this such as reformulating the problem, warping or employing distribution independent methods.

Since there is no ground truth data for the passenger delay associated with each network element and the estimation framework are based on inferred variables, the estimates might be biased. The biases can stem from either the inferred trajectory resulting from the Passenger-to-Train assignment or from the least square approach itself. The in-vehicle time and transfer time would not be affected by these biases as the system of equations are also constrained by the AVL data (essentially serve as the ground truth data for these estimates). As for the bias from least squares, it can mainly occur when the solution is not unique. This can happen when data is insufficient, but this bias can be avoided (e.g. time slices definition) and relates to the robustness of the approach.

#### 4. Conclusion

In this study, we propose a new estimation method to map passenger delay into network elements. The outputs of our method can aid in measuring network performance for any given origin-destination pair of the public transportation network and in prioritizing measures for improving service robustness. We decompose the delay along a passenger trajectory into its corresponding track segment delay, initial waiting time and transfer delay. We demonstrate the method using one-year data from the Washington metro network.

Our method estimates how passenger delays are distributed across network elements and can thus assess the contribution of each network element to system performance. We were able to achieve a dimensionality reduction of 94% by representing these individual trajectories as 3D networks. The estimation results show that the confidence of the estimation, measured based on the compactness of the boxplot, increases with an increase in the number of passengers. Overall, our estimation framework achieved a mean slack of 0.072 min or less than 5 s per passenger, which is very promising.

These network element estimates are used to analyze the average and total passenger delay of the metro network. The average passenger delay is more or less stable throughout the day whereas total passenger delay, which accounts for the number of passengers affected, contains two distinct peaks. The two peaks correspond to the morning peak and evening peak with the evening peak associated with greater passenger delay than the morning peak. The initial waiting time contributes the most to the average passenger delay with 59% whereas the track segment delay contributes the most to the total passenger delay with 41% of the total delay being attributed to it.

The estimates generated by the method proposed in this paper open new avenues for future research. In particular, a more sophisticated distribution independent method can be used to solve the system of equations instead of the least square approach. Furthermore, better error assessment for each individual network element estimates can be done using additional personal data such as disaggregated personal travel diaries and video data. The estimation framework can also be adapted to allow non-uniform temporal aggregation to reduce the share of passenger delay records for which the slack variable value is negative. The delay estimates at the individual network element level can be used to reveal hourly, daily, weekly or seasonal delay patterns across the metro network. Moreover, the estimation approach is easily transferable to other locations. For this, it is necessary to ensure that the system of equations is solvable and that a different temporal aggregation might be needed depending on the headway between the transit vehicles. Furthermore, it can help in understanding the delay propagation through a network and potentially contribute to the prediction of such delays and their spill-over impacts. Such advancements can ultimately help in supporting decision making in relation to improved service robustness at both the tactical level (e.g. locating switches to allow for short-turning) and operational level (e.g. disruption mitigation strategies including information provision and resource allocation).

#### Acknowledgements

This research was supported by the SETA project which is financed by the European Union's Horizon 2020 Research and Innovation program under the grant agreement No 688082. The authors would like to thank Washington Metropolitan Area Transit Authority and in particular Jordan Holt for their valuable cooperation and providing the data that made this study possible. We thank the anonymous reviewers for the constructive feedback which has greatly improved this paper.

## Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at <https://doi.org/10.1016/j.trc.2020.102704>.

## References

- Altman, A., Gondzio, J., 1999. Regularized symmetric indefinite systems in interior point methods for linear and quadratic optimization. *Optim. Methods Software* 11 (1–4), 275–302.
- Bhaskar, A., Chung, E., et al., 2014. Passenger segmentation using smart card data. *IEEE Trans. Intell. Transport. Syst.* 16 (3), 1537–1548.
- Box, G.E., 1976. Science and statistics. *J. Am. Stat. Assoc.* 71 (356), 791–799.
- Gordon, J.B., Koutsopoulos, H.N., Wilson, N.H., Attanucci, J.P., 2013. Automated inference of linked transit journeys in london using fare-transaction and vehicle location data. *Transp. Res. Rec.* 2343 (1), 17–24.
- Hayashi, F., *Econometrics*. 2000, Princeton University Press. Section, 1, pp. 60–69.
- Hendren, P., Antos, J., Carney, Y., Harcum, R., 2015. Transit travel time reliability: shifting the focus from vehicles to customers. *Transp. Res. Rec.* 2535 (1), 35–44.
- Koutsopoulos, H.N., Noursalehi, P., Zhu, Y., Wilson, N.H., 2017. Automated data in transit: Recent developments and applications. In: 2017 5th IEEE international conference on models and technologies for intelligent transportation systems (MT-ITS), IEEE, 2017, pp. 604–609.
- Krishnakumari, P., van Lint, H., Djukic, T., Cats, O. A data driven method for od matrix estimation. *Transport. Res. Part C: Emerg. Technol.* doi:10.1016/j.trc.2019.05.014.
- Ma, X., Wu, Y.-J., Wang, Y., Chen, F., Liu, J., 2013. Mining smart card data for transit riders' travel patterns. *Transport. Res. Part C: Emerg. Technol.* 36, 1–12.
- Ma, Z.-L., Ferreira, L., Mesbah, M., Hojati, A.T., 2015. Modeling bus travel time reliability with supply and demand data from automatic vehicle location and smart card systems. *Transp. Res. Rec.* 2533 (1), 17–27.
- Ma, X., Liu, C., Wen, H., Wang, Y., Wu, Y.-J., 2017. Understanding commuting patterns using transit smart card data. *J. Transp. Geogr.* 58, 135–145.
- Ma, Z., Zhu, S., Koutsopoulos, H.N., Ferreira, L., 2017. Quantile regression analysis of transit travel time reliability with automatic vehicle location and farecard data. *Transp. Res. Rec.* 2652 (1), 19–29.
- Moreira-Matias, L., Mendes-Moreira, J., de Sousa, J.F., Gama, J., 2015. Improving mass transit operations by using avl-based systems: A survey. *IEEE Trans. Intell. Transp. Syst.* 16 (4), 1636–1653.
- Nassir, N., Khani, A., Lee, S.G., Noh, H., Hickman, M., 2011. Transit stop-level origin–destination estimation through use of transit schedule and automated data collection system. *Transport. Res. Rec.* 2263 (1), 140–150.
- Pelletier, M.-P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: A literature review. *Transport. Res. Part C: Emerg. Technol.* 19 (4), 557–568.
- Sánchez-Martínez, G.E., 2017. Inference of public transportation trip destinations by using fare transaction and vehicle location data: Dynamic programming approach. *Transp. Res. Rec.* 2652 (1), 1–7.
- Trépanier, M., Tranchant, N., Chapleau, R., 2007. Individual trip destination estimation in a transit smart card automated fare collection system. *J. Intell. Transport. Syst.* 11 (1), 1–14.
- Trépanier, M., Morency, C., Agard, B., 2009. Calculation of transit performance measures using smartcard data. *J. Public Transport.* 12 (1), 5.
- Yap, M., Cats, O., Van Oort, N., Hoogendoorn, S., 2017. A robust transfer inference algorithm for public transport journeys during disruptions. *Transport. Res. Proc.* 27, 1042–1049.
- Yap, M., Cats, O., van Arem, B., 2018. Crowding valuation in urban tram and bus transportation based on smart card data. *Transportmet. A: Transp. Sci.*, pp. 1–20.
- Zhao, J., Frumin, M., Wilson, N., Zhao, Z., 2013. Unified estimator for excess journey time under heterogeneous passenger incidence behavior using smartcard data. *Transport. Res. Part C: Emerg. Technol.* 34, 70–88.
- Zhao, J., Zhang, F., Tu, L., Xu, C., Shen, D., Tian, C., Li, X.-Y., Li, Z., 2016. Estimation of passenger route choice pattern using smart card data for complex metro systems. *IEEE Trans. Intell. Transp. Syst.* 18 (4), 790–801.
- Zhu, Y., Koutsopoulos, H.N., Wilson, N.H., 2017. A probabilistic passenger-to-train assignment model based on automated data. *Transport. Res. Part B: Methodol.* 104, 522–542.