



# User-based representation of time-resolved multimodal public transportation networks

Laura L Alessandretti, Márton Karsai, Laetitia Gauvin

## ► To cite this version:

Laura L Alessandretti, Márton Karsai, Laetitia Gauvin. User-based representation of time-resolved multimodal public transportation networks. Royal Society Open Science, 2016, 3, pp.160156. 10.1098/rsos.160156 . hal-01249860

**HAL Id: hal-01249860**

**<https://inria.hal.science/hal-01249860v1>**

Submitted on 5 Jun 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



**Cite this article:** Alessandretti L, Karsai M, Gauvin L. 2016 User-based representation of time-resolved multimodal public transportation networks. *R. Soc. open sci.* **3**: 160156.  
<http://dx.doi.org/10.1098/rsos.160156>

Received: 3 March 2016

Accepted: 17 June 2016

**Subject Category:**

Computer science

**Subject Areas:**

environmental science/civil  
engineering/pattern recognition

**Keywords:**

public transportation, multimodal networks,  
human dynamics

**Author for correspondence:**

Márton Karsai

e-mail: [marton.karsai@ens-lyon.fr](mailto:marton.karsai@ens-lyon.fr)

One contribution to a special feature 'City analytics: mathematical modelling and computational analytics for urban behaviour'.

Electronic supplementary material is available at <http://dx.doi.org/10.1098/rsos.160156> or via <http://rsos.royalsocietypublishing.org>.


# User-based representation of time-resolved multimodal public transportation networks

Laura Alessandretti<sup>1,2,3</sup>, Márton Karsai<sup>1</sup> and  
Laetitia Gauvin<sup>2</sup>

<sup>1</sup>Université de Lyon, ENS de Lyon, LIP, INRIA-CNRS-UMR 5668, IXXI, 69364 Lyon, France

<sup>2</sup>Data Science Lab, ISI Foundation, Turin, Italy

<sup>3</sup>Department of Mathematics, City University London, London EC1V 0HB, UK

 MK, 0000-0001-5752-556X

Multimodal transportation systems, with several coexisting services like bus, tram and metro, can be represented as time-resolved multilayer networks where the different transportation modes connecting the same set of nodes are associated with distinct network layers. Their quantitative description became possible recently due to openly accessible datasets describing the geo-localized transportation dynamics of large urban areas. Advancements call for novel analytics, which combines earlier established methods and exploits the inherent complexity of the data. Here, we provide a novel user-based representation of public transportation systems, which combines representations, accounting for the presence of multiple lines and reducing the effect of spatial embeddedness, while considering the total travel time, its variability across the schedule, and taking into account the number of transfers necessary. After the adjustment of earlier techniques to the novel representation framework, we analyse the public transportation systems of several French municipal areas and identify hidden patterns of privileged connections. Furthermore, we study their efficiency as compared to the commuting flow. The proposed representation could help to enhance resilience of local transportation systems to provide better design policies for future developments.

## 1. Introduction

Urban transportation systems interweave our everyday life and although their construction is based on conscious design they appear with complex structural and dynamical features [1]. They build up from different transportation means, which connect

places in a geographical space. Their most straightforward description is given by networks [2,3] where stations are identified as nodes, and links are the transportation connections between them. Based on this representation [4], considerable research efforts have been dedicated to address their sustainability [5], to optimize their efficiency [6,7], reliability [8–10] or even to estimate the risk they carry due to interdependency with other infrastructure networks in case of terrorist attacks [11].

All transportation networks share a few common features: (i) they are all embedded in space, setting constraints in their structural design, (ii) networks of different transportation means may coexist in the same space and (iii) they are all inherently temporally resolved. Such details of several transportation networks became available lately [12] through the collection of large open datasets describing complete multimodal transportation systems in cities, regions, countries and even internationally. These advancements were induced by novel data collection techniques including smart card data [13], automatic vehicle location (AVL) data [14] and mobile phone data from GSM providers [15]. On the top of these developments, the advent of a new common non-proprietary transit data format, the General Transit Feed Specification (GTFS), further amplified actual trends in urban policy propagating smart city programmes and real-time online user services. As of February 2016, 325 public transportation companies around the world have released official GTFS feeds [16], which are regularly modified by online communities that are adding extensions and optional fields to adapt to different transit services [17]. In transportation, the confluence of open data, GTFS, ubiquitous mobile computing, sensing and communication technologies, has allowed to study the efficiency and performance of public transportation systems under different perspectives [18–22]. As a consequence, GTFS data are now used for trip planning, ride-sharing, timetable creation, mobile data, visualization, accessibility and to provide real-time service informations.

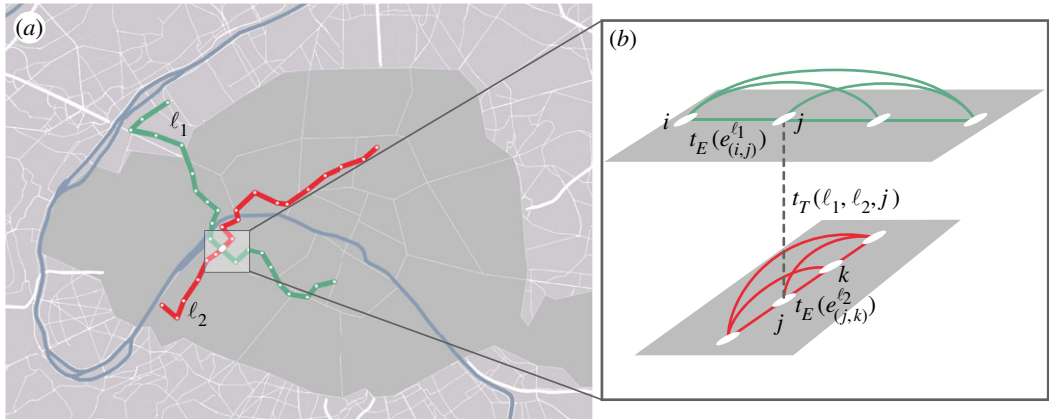
These recent developments in data collection practices and in the corresponding fields of complex networks and human dynamics provided the opportunity to quantitatively study transportation systems using a data-driven approach. These studies showed that geographical constraints largely determine the structure and scaling of transportation networks [23–25], but for their better understanding one needs to consider the actual urban environment and development level [7,26,27]. At the same time, the emerging field of multilayer networks provided the methodology to consider their multimodal character [28,29]. In this representation, each layer corresponds to the network of a single means of transportation (bus, tram, train, etc.), which are defined on the same set of nodes (stations). This way they account for possible multiple links of different modes between the same stations [30]. This representation can be extended to capture the temporal nature of the system by using some aggregated information extracted from the transportation schedule [31] or, as a future challenge, by considering each time slot as a layer where journeys between stations are represented as temporal links [28,32].

Here, we build on these contemporary advancements and provide a novel representation, which combines multi-edge and P-space representations of transportation networks. The proposed scheme considers the system from the user's point of view by incorporating the minimization of the total travel time, its variability across the schedule, and the number of transfers between lines. Our subsequent aim is to adjust earlier-defined characterization techniques to the proposed representation in order to help its analysis. We use the adjusted techniques (i) to identify patterns of privileged connections in the transportation network, which are not evidently present through their overall design, and (ii) to quantify their overall efficiency when compared with the commuting flow. We carried out our analysis using openly shared GTFS datasets describing extensive transportation networks of French municipalities, such as greater Paris, Toulouse, Nantes and Strasbourg.

As follows, first, we describe the actual time-resolved multilayer network representation and introduce our methodology incorporating travel routes and times to identify efficient transportation connections. Next, we apply a matrix factorization method to extract underlying connectivity patterns to analyse them from the commuter point of view, and quantify their overall efficiency. Finally, we conclude our results and discuss possible applications and future directions of research. Note that the implementation of the proposed methodology is openly accessible online ([https://github.com/lalessan/user\\_basedPT](https://github.com/lalessan/user_basedPT)).

## 2. Representation of public transportation networks

The proposed methodology integrates several sequential steps to detect origin–destination areas that are conveniently connected by public transportation with respect to user preferences. In the following description, first, we define a user-based representation of a public transportation (PT) system, which



**Figure 1.** Illustration of the user-based multi-edge P-space representation. (a) Two geo-localized crossing PT lines  $\ell_1$  and  $\ell_2$  are shown on the map of central Paris. (b) Schematic of the P-space multi-edge representation for a section of the network: all pairs of nodes corresponding to stops on the same line are connected by edges with the same label.

limits the effect of its spatial embeddedness, but accounts for its multilayer structure, and its temporal dimension. Next, we calculate the shortest time paths between stops by adapting a conventional algorithm [33] to the actual graph representation, and finally, we select preferred connections, taking into account the distance travelled over time.

## 2.1. User-based multi-edge P-space representation

Earlier studies revealed that the choice of users to select transportation means for commuting is mainly affected by the average travel time, and by the variability of the total travel time [34],<sup>1</sup> [35],<sup>2</sup> in addition to the number of transfers they need to do. Our principal goal here is to introduce a novel representation of PT networks, which incorporates the aforementioned aspects decisive for users (while neglecting other presumably less determinant factors such as travel cost or comfort), and which minimizes the effects due to the spatial embeddedness of the system. In order to do so, we combine a multi-edge [36] and a P-space representation of the transportation network [37–39] to describe the PT systems. The multi-edge representation accounts for the presence of several transportation lines in the same PT network by allowing the existence of multiple labelled edges within a single pair of nodes. On the other hand, the P-space representation takes into account that transfers between lines is time-consuming and may not be convenient for the user; also, it considers connections between stops located at large distance thus it reduces the effect of the geographical distances. The combination of these two representations constitutes an ideal framework to investigate complex features of PT systems from the user perspective. A schematic example of this representation is displayed in figure 1a, showing two crossing PT lines, and in figure 1b, we illustrate the corresponding P-space multi-edge representation. Two stops  $i$  (respectively,  $k$ ) and  $j$  on the same line  $\ell_1$  (respectively,  $\ell_2$ ) are linked through the edge  $e_{(i,j)}^{\ell_1}$  (respectively,  $e_{(j,k)}^{\ell_2}$ ), with weight  $t_E(e_{(i,j)}^{\ell_1})$  (respectively,  $t_E(e_{(j,k)}^{\ell_2})$ ). At node  $j$ , a transfer is possible between the two lines, which is represented by a link with weight  $t_T(\ell_1, \ell_2, j)$  corresponding to the actual time of transfer.

Formally, the public transportation system is defined as a weighted, directed, edge-labelled graph  $G = (V, E, t_E, T, t_T)$  with vertex set  $V$  with cardinality  $N$ , corresponding to the public transportation stops, edge set  $E$  with weight function  $t_E$ , and set of transfers  $T$  with weight function  $t_T$ . If a line  $\ell_k$  is defined as an ordered sequence of stops connected consecutively, in the corresponding P-space graph  $G$ , there will be a direct labelled-edge  $e_{ij}^{\ell_k} \in E$  connecting each pair of nodes  $(i, j)$  on the given line, such that stop  $i$  precedes stop  $j$  in the sequence of line  $\ell_k$ . This way each transportation line appears as a fully connected clique in the P-space representation. We define  $M$  as the total number of lines in the PT system. Furthermore, a set  $T \subset M \times M \times N$  of transfers identifies triplets of two lines and one node,

<sup>1</sup>This document provides a study on the factors influencing the choice of the transportation means. The study is based on a literature review and statistical analysis of surveys.

<sup>2</sup>This document presents results of a survey about transport and mobility of households in France. The authors consider the distance travelled, if it is a long-distance trip or short trip made on a daily basis. Moreover, they look at the differences of behaviours in transportation use among different regions and individuals with different socio-demographic characteristics.

$e_{\ell_1, \ell_2, j}^T = (\ell_1, \ell_2, j)$  assigning a possible transfer between lines  $\ell_1$  and  $\ell_2$  at station  $j$ . Each edge in  $E$  is weighted by the average travel time on the actual line. It is computed through a *time* function  $t_E : E \rightarrow \mathbb{R}^+$ , quantifying for each edge  $e_{ij}^{\ell_k}$  the time needed to get from  $i$  to  $j$  along the line  $\ell_k$  averaged on a selected time window  $[h1, h2]$  over  $N_w$  weeks. The travel time assigned to an edge  $e_{ij}^{\ell_k} \in E$  is then calculated as the sum of the average waiting time and the average time spent on the vehicle as

$$t_E(e_{ij}^{\ell_k}) = \frac{1}{2f_{\ell_k}} + \Delta t_{ij}^{\ell_k}, \quad (2.1)$$

where  $f_{\ell_k}$  is the average frequency of line  $\ell_k$  and  $\Delta t_{ij}^{\ell_k}$  is the average time one needs to spend on line  $\ell_k$  to go from stop  $i$  to stop  $j$ . This formula is designed to consider the case where a user would go blindly to a stop (without looking at the schedule). Another approach would include that certain passengers attempt to reduce their waiting time by timing their arrival at transit stops to an optimal period before vehicle departure. Most studies report that passengers facing short headways or low reliability do not generally pursue these strategies [40–42]. Hence, we choose our approach to favour lines with high frequency and less variability due to unexpected perturbations while accounting for preference of users for low unexpected variability in the total travel time. Finally, the transfer time function  $t_T : T \rightarrow \mathbb{R}^+$  quantifies for each transfer  $e_{\ell_1, \ell_2, j}^T$  the time needed to change between lines  $\ell_1$  and  $\ell_2$  at node  $j$ .

In such description, the temporality of the system is included through the weights. The choice not to model the system as a temporal graph is motivated by the fact that in urban public transportation systems the total travel time is subject to variability and this factor matters considerably for the user when deciding to opt for public transportation service.

## 2.2. Uncovering efficient transportation connections

The previously defined public transportation graph  $G = (V, E, t_E, T, t_T)$  is used to calculate the shortest time paths between stops. In the multi-edge representation a path is defined as a sequence of edges<sup>3</sup>  $P_E = \{e^{\ell_{i_1}}, e^{\ell_{i_2}}, \dots, e^{\ell_{i_n}}\}_{o,d}$  connecting an origin node  $o$  to a destination node  $d$  through a sequence of consecutive trips made on  $n$  lines,  $\ell_{i_1}, \ell_{i_2}, \dots, \ell_{i_n}$ . Considering also the sequence of corresponding transfers between lines  $P_T = \{e_{\ell_{i_1}, \ell_{i_2}}^T, e_{\ell_{i_2}, \ell_{i_3}}^T, \dots, e_{\ell_{i_{n-1}}, \ell_{i_n}}^T\}_{o,d}$  the shortest time paths between origin and destination are taken as the smallest durations measured among the different alternative paths. Each time length is defined as

$$L_P = \sum_{j=1}^n t_E(e^{\ell_{i_j}}) + \sum_{j=1}^{n-1} t_T(e_{\ell_{i_j}, \ell_{i_{j+1}}}^T), \quad (2.2)$$

i.e. the sum of the average time needed to wait, travel and transfer between lines.

We adapted the Dijkstra algorithm [33] to provide approximated shortest path lengths between any pair of stops in the user-based multilayer representation, while keeping the interpretable description of the PT system and reduced computation time. The original version of the algorithm computes the minimal distance between any origin  $o$  and destination  $d$  nodes by considering the sum of link weights. Instead, the modified version accounts for the fact that not only the link weights have to be taken into consideration, but also the transfer time, i.e. the cost to change between different layers (see the electronic supplementary material, section S3 for more details). Also, to consider the preference of users to change lines a limited number of times, the algorithm allows at most two transfers in a single path, i.e. we limit  $n \leq 3$ . Owing to these limitations, the algorithm provides us an approximate solution, which, however, differs from the correct solution in only few cases. We find that more than 95% of the paths with at most three line changes computed with the unlimited (correct) algorithm and the limited (approximate) algorithm have the same temporal length in all cities (see the electronic supplementary material). After computing the shortest paths between all nodes in the graph, we characterize the distribution of shortest travelling times between all nodes whose physical distance falls within a specific range. Using this information, we identify privileged connections, i.e. fastest routes at a given distance.

## 2.3. Implementation of the user-based representation

The methodology presented above relies on information, which is typically included in data given in GTFS format (<https://developers.google.com/transit/gtfs/reference>) such as trips, routes, travelling times, frequencies and transfer times recorded for each service line and station in the transportation

<sup>3</sup>In the current paragraph, to simplify notations, we do not index edges by node names.



system (for further details, see the electronic supplementary material, section S1). Using such data, we build the P-space multi-edge representations of greater Paris, Strasbourg, Nantes and Toulouse. We decided to use a period of  $N_w = 4$  weeks in each case, such that the total number of trips per day presents only weak fluctuations. We were interested in trips planned between  $h1 = 7.00$  h and  $h2 = 10.00$  h (though the choices of  $N_w$ ,  $h1$ , and  $h2$  are adjustable parameters). This choice of time window was made to focus on morning commuting patterns, and because during this time interval, the frequency of services is considerably higher than for the rest of the day. Typical line frequencies and trip durations are then defined as their averages over the selected time window over the four weeks. All PT systems considered rely substantially on three transportation modalities: metro (Paris, Toulouse) or tram (Nantes and Strasbourg), bus and rail. However, they differ considerably in terms of size (see the electronic supplementary material, table S2), route length, number of stops per route and route frequencies (see the electronic supplementary material, figure S1).

Finally, building on the multi-edge P-space representation and the estimation of the typical times and frequencies, we compute the typical shortest time paths between pairs of origin and destination in the city. The implementation of this methodology is available online ([https://github.com/lalessan/user\\_basedPT](https://github.com/lalessan/user_basedPT)) and requires as input any dataset in GTFS format and parameters summarized in the electronic supplementary material.

### 3. Illustration: fingerprints of public transportation networks

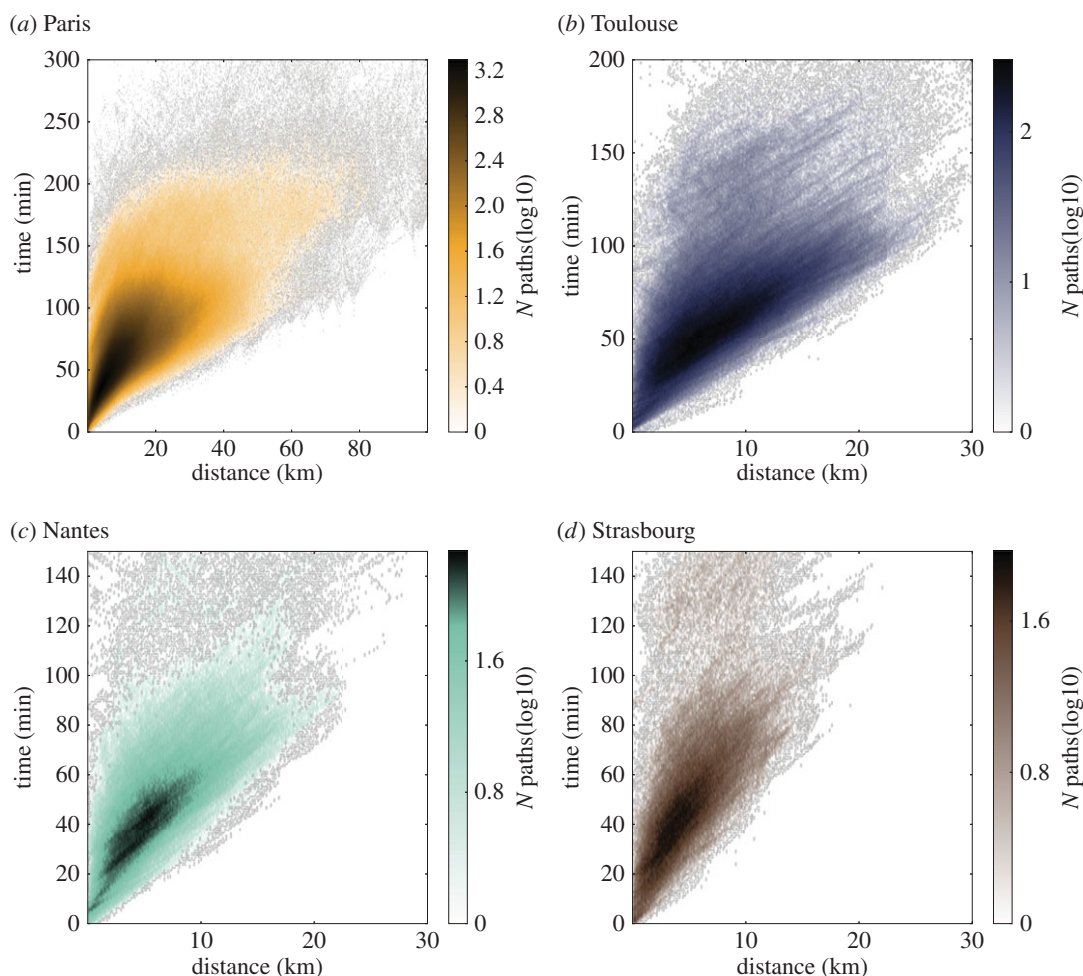
We demonstrate one possible use of our framework through the examples of the PT systems of greater Paris, Strasbourg, Nantes and Toulouse. After selecting privileged connections, we apply non-negative matrix factorization (NMF) to the graph of the privileged connections to identify underlying patterns, which may not be present due to overall design. Finally, we compare our findings with independent measures of commuting patterns, which allow us to give an estimation of the efficiency of the PT systems.

#### 3.1. Selection of efficient connections

We used the method previously presented to compute the shortest time paths for all origin–destination pairs of the transportation systems of bus, train and metro. With the selection of a suitable time-window of size  $N_w = 4$  weeks, we find that the average standard deviation of a route frequency across the 4 weeks is about 0.05 transits/hour for all the cities considered (focusing only on weekdays). This result confirms that the system behaviour is subject to low variability during the period considered. Based on the shortest path calculations, we built a time–distance map, which assigns the physical distance  $d(o, d)$  and the shortest time path length  $\Delta t(o, d)$  to each origin ( $o$ )–destination ( $d$ ) pair. This time–distance map was drawn as a heat-map in figure 2 for Paris and the other cities investigated, and can be used to identify patterns of privileged connections. We considered distance-bins with equal size 100 m and time bins of size 1 min.

In order to focus on the most efficient (privileged) connections with respect to the public transportation system of the city considered, we selected the trips responsible for the lowest 1% of the time distributions for each distance. To estimate whether these connections are among the best at the urban agglomeration level when compared with travel by car for the same distances, we computed the travel time factor. More precisely, after building the histogram of the shortest time paths for every distance bin, we compared the travel time of selected paths with the travel time needed to cover the same distance by car. Car commuting times were extracted from the French 2008 Enquête Nationale Transports et Déplacements 2007–2008 dataset [43]<sup>4</sup> describing the global mobility of people living in France. To collect these data, individuals were asked how far (with resolution of 1 km), how long (with resolution of 1 min) and by which means of transportation they travel every day. Based on this dataset, we computed the median of the travel time distributions at each distance using the entire sample to measure the typical time needed to commute a particular distance by car. Similarly, we calculated the medians of the best 1, 2 and 5% of the time distribution at each distance (i.e. shortest times for a given distance) travelled by public transportation. This enables to compute the travel time factor as displayed in figure 3 for different selections of the best times taken by public transportation. By selecting the best connections responsible for the lowest 1% of the time distributions for each distance, in Paris agglomeration, we found that trip durations are at most 1.71 times the time needed by car. This is in close agreement with the travel time factor tolerated by users [34], which was shown to be maximum 1.6 in [34]. For the other

<sup>4</sup>This reference links to a file about the home–work flow of individuals by transportation means.



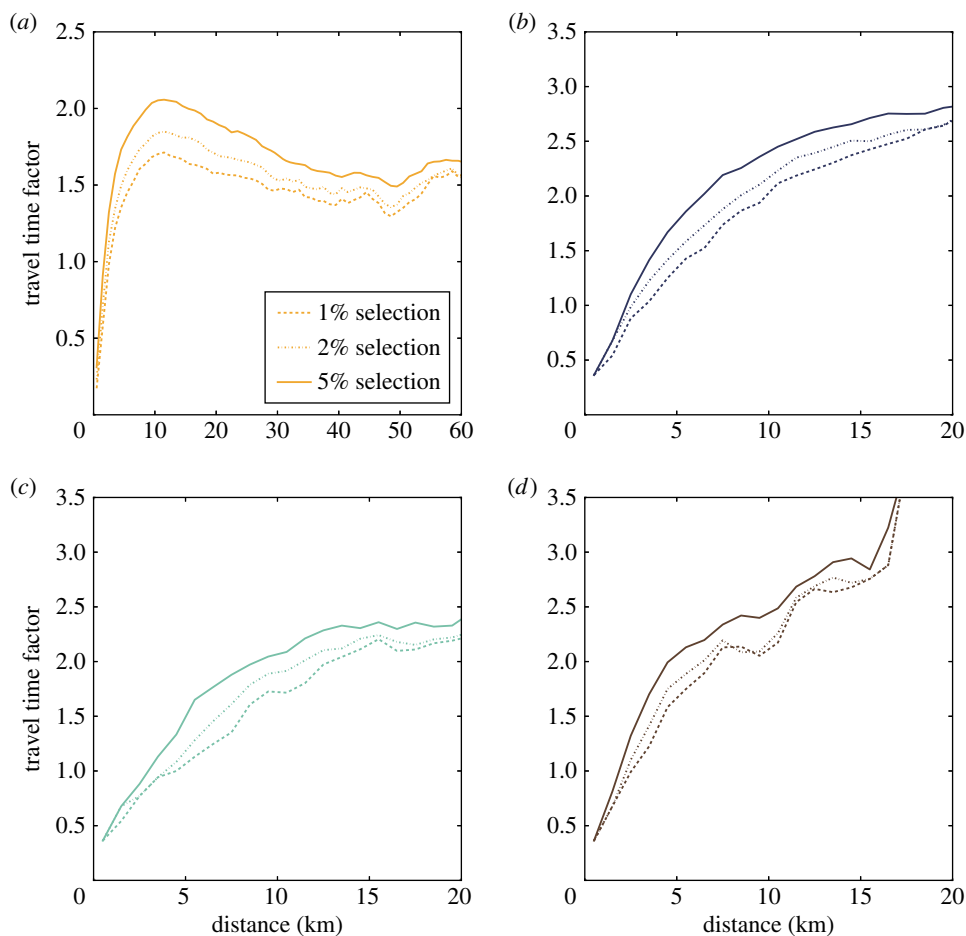
**Figure 2.** Scatterplot of time versus physical distance associated with the shortest time paths for each origin–destination pair. The points are coloured according to the number of points in the area considered. Scatterplots are shown for the cities of (a) Paris, (b) Toulouse, (c) Nantes and (d) Strasbourg. Colours indicate the logarithm of the number of origin–destination pairs in a given range time–distance bin.

agglomerations studied, the travel time factor goes above this value for distances travelled greater than 5 km. We note that while in Paris the travel time factor tends to saturate at large distance, meaning that efficient connections exist also at the inter-city level, this is not the case for the other cities (figure 3*b–d*), where PT seems to provide an efficient alternative to car mainly for short trips.

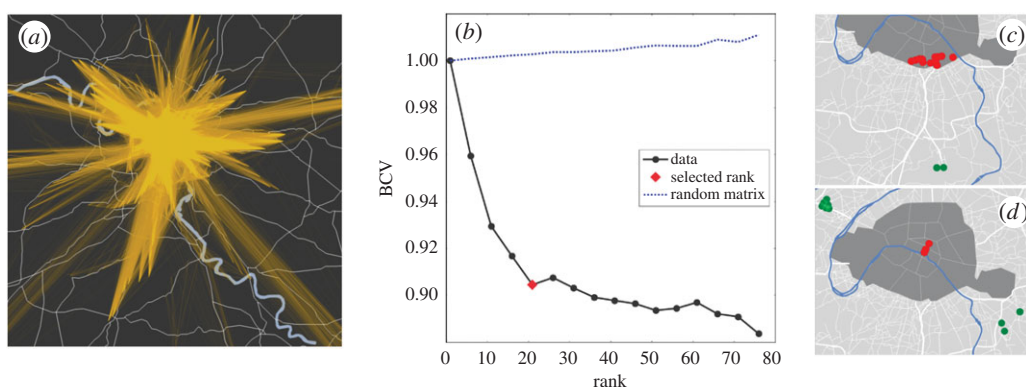
Let us notice again that in the histograms and travel time factor calculations, we do not use the best absolute time to travel a given distance but we consider a waiting time assuming that a user arrives blindly at a stop, in order to take into account the preferences of users for paths with small variability in time. In addition, the time travelled by car for each distance is taken from data considering car trips in the whole country. These two points may lead to an overestimation of time travel factors, so the travel time factor cannot be used directly as a criterion to select the best connections but only gives a common metric to look at the different public transportation systems.

For all the cities considered, privileged connections include shortest paths with no line changes at very short distances, and an increasing number of line changes as a function of the distance travelled (see the electronic supplementary material, figure S6). In the Paris area, only two transportation modalities are used in 80% of the paths up to 60 km distance and the most-represented modalities include metro, bus and rail, with metro dominating at short distances. In smaller cities, almost all paths at distances up to 20 km involve less than two changes. The most-represented modalities are bus and tram for Strasbourg and Nantes, bus and metro for Toulouse; instead, rail is present only when the path distance is larger than 15/20 km (see the electronic supplementary material, figures S7 and S8).

In figure 4*a*, we show the profile of the Paris urban agglomeration, where links correspond to the selected privileged connections. The city profile differs clearly from the profile obtained for



**Figure 3.** Travel time factors with respect to distance travelled. The factors have been computed using the lowest 1, 2 and 5% of the time distribution for each distance travelled by public transportation for the following cities (including their surrounding areas) (a) Paris, (b) Toulouse, (c) Nantes and (d) Strasbourg.



**Figure 4.** Pattern detection using the multi-edge P-space representation. (a) Geographical representation of graph  $G_{sp}$ , where links correspond to the 1% best shortest paths of the whole public transportation network. (b) The normalized BiCross validation error computed for the adjacency matrix  $X_{sp}$  (10 km, 11 km) (solid black line) of the same graph, for the associated random matrix  $X_{sp,random}$  (10 km, 11 km) (dotted line). The selected number of structures  $k_s$  is assigned by a red rhombus. (c, d) Two of the structures revealed in the PT system of Paris. Green dots are ingoing, while red dots are outgoing affiliated.

single-modality single-layer representations, since it accounts for the interconnectedness of several transportation modes (see the electronic supplementary material, section S7). Note that, since we do not have access to the transfer times between lines in cities other than in Paris, the cost of transferring



between layers was estimated for each city based on the data for Paris (see the electronic supplementary material, table S3). This way, transfer times depend on the corresponding transportation modes, which a naive representation with all modes on a single layer would not be able to consider.

## 3.2. Pattern extraction

The question remains whether the identified set of privileged connections reveals any higher-order meaningful patterns in the design of transportation systems. We expect that some stops, such as stations located in residential neighbourhoods, may have similar connectivity patterns to the rest of the network as to the city centre or to working areas. In order to identify such patterns, we first built an undirected, unweighted graph  $G_{SP} = (V_{SP}, E_{SP})$ , where  $V_{SP} \subset V$  and  $E_{SP}$  are a set of edges linking origin–destination locations connected by privileged connections (for an example for Paris, see figure 4a). To compare commuters travelling at particular distances, we analysed subgraphs  $G_{SP}(d_1, d_2)$  (represented by an adjacency matrix  $X_{SP}(d_1, d_2)$ ) of  $G_{SP}$ , where edges join stops at particular distances  $d$  ( $d_1 < d \leq d_2$ ). For Paris, we considered distances with resolution  $d_2 - d_1 = 1$  km, while for smaller cities we took the resolution  $d_2 - d_1 = 5$  km as the transportation networks were typically sparser there (see the electronic supplementary material, figure S1).

We expected to find both cohesive and bipartite patterns in these subgraphs. The cohesive structures would correspond to sets of stations well connected among themselves, while bipartite ones would single out two groups of stops with several connections between them. The connections may not be direct but should have durations comparable to the average time taken by car for the same distance.

To detect such patterns, we considered the likelihood of having a connection between any two stations, which can be expressed in terms of possible connections of these stations to the same structures. Formally, it means we can express each term of the adjacency matrix representing  $G_{SP}$  as

$$X_{SP}(i, j) = \sum_k W_{ik} H_{kj}, \quad (3.1)$$

where  $W_{ik}$  quantifies the ingoing membership of node  $i$  to structure  $k$  and  $H_{kj}$  quantifies the outgoing affiliation of the node  $j$  to the structure. In order to find matrices  $\mathbf{W}$  and  $\mathbf{H}$ , we performed matrix factorization, thus minimizing numerically the distance

$$\|\mathbf{X} - \mathbf{WH}\|_F^2, \quad (3.2)$$

where  $\|\mathbf{X}\|_F$  is the Frobenius norm of matrix  $\mathbf{X}$  (for further details see the electronic supplementary material, section S2). Note that matrix factorization was used earlier successfully to detect communities and higher-order structures in graphs [44–50].

The number of structures to be detected was determined by the Bi-Cross validation (BiCv) approach proposed in [51] based on cross-validation, a common machine learning model validation technique. This consists of measuring an error, called BCV here, between an estimation of left-out entries using a low-rank approximation of the retained data and the actual left-out entries. This error is decreasing with respect to the number of structures extracted towards a minimum that indicates how many structures are representative of the subgraphs (more details in the electronic supplementary material, section S4), while, on the contrary, such behaviour is not visible when the network is close to random. To identify whether there are structures in subgraphs, we compared the BCV error with the one obtained for the corresponding null models (figure 4b). Such null models were defined for each adjacency matrix  $X_{SP}(d_1, d_2)$  as their corresponding random matrices  $X_{SPrandom}(d_1, d_2)$  with the same size and density. An example of the behaviour of such a quantity for the Paris public transportation network is displayed in figure 4b (for other cities, see the electronic supplementary material, figure S1). This quantity was computed for each subgraph and guided us on how many structures characterize each system at each range of distance. For some distance ranges and cities, the evolution of BCV is close to the random case assigning no strong attempt to link preferentially some areas at the considered range of distance (see the electronic supplementary material, figure S2). However, we find bipartite structures in several cases, like the two examples in figure 4c,d detected in the Paris network. The bipartite structures can be assimilated to strategical areas that are particularly well connected by PT. For example, the structure shown in figure 4c connects stops located around Paris Orly airport to stops located at the border of the Paris central area. In figure 4d, the structure reveals the existence of privileged connections between the Nanterre and Creteil areas on one side (both with high employment density; [http://insee.fr/fr/themes/document.asp?reg\\_id=20&ref\\_id=20718&page=alapage/alap417/](http://insee.fr/fr/themes/document.asp?reg_id=20&ref_id=20718&page=alapage/alap417/)

[alap417\\_carte.htm#carte1](#)) and Paris centre on the other side. As these structures are latent patterns extracted from the networks of privileged connections, we consider them as the privileged origin–destination patterns representative of the transportation systems.

### 3.3. Network efficiency: pattern analysis from the commuter point of view

To estimate how well the different public transportation networks are devoted to answer the needs of commuters, we compared the identified privileged origin–destination patterns to the flows of commuters. We used the data of the 2010 French census [52] including origin–destination commuter flows per means of transportation at the level of the municipality for the areas of greater Strasbourg, Toulouse and Nantes, and at the level of the municipal arrondissement (neighbourhood) for the Paris agglomeration. Using this dataset, we compared the detected privileged origin–destination patterns to the commuting patterns by car and PT. We only considered inter-municipality trips for the comparison as the resolution provided for the commuter dataset was given at the municipality level (for the number of intra-city trips, see the electronic supplementary material, table S4).

To draw a comparison, we first built the PT structural pattern network  $G_C = (V_C, E_C)$  of each urban agglomeration as an unweighted, undirected graph. Here, the set of nodes  $V_C$  is defined as municipalities and a link  $(a, b) \in E_C$  between municipalities  $a$  and  $b$  exists if at least one stop located in  $a$  and one stop in located  $b$  appear in each side of a detected bipartite structure. In other words, the structural pattern networks are composed of links between municipalities presumably well connected by public transportation. At the same time, exploiting census data, we built a commuter flow network for each city and its surrounding area, as a weighted, directed graph  $G_{\text{com}}^{TM} = (V_{\text{com}}^M, E_{\text{com}}^{TM}, W_{\text{com}}^M)$ . Here,  $V_{\text{com}}^M$  is the set of municipalities, and a link  $(a, b) \in E_{\text{com}}^M$  with weight  $w_{ab}$  represents the flow of individuals commuting from  $a$  to  $b$  by means  $TM$  (either PT or car). We compared the structural pattern graph with the commuter flow graphs both of the car and the PT of each urban agglomeration by computing a weighted Jaccard index  $s$  between the sets of links associated with each graph. This weighted index is defined as the sum of the flow graph weights of the links in common between the two graphs—structural and flow by the selected transportation means—divided by the total flow for the transportation means considered. More formally

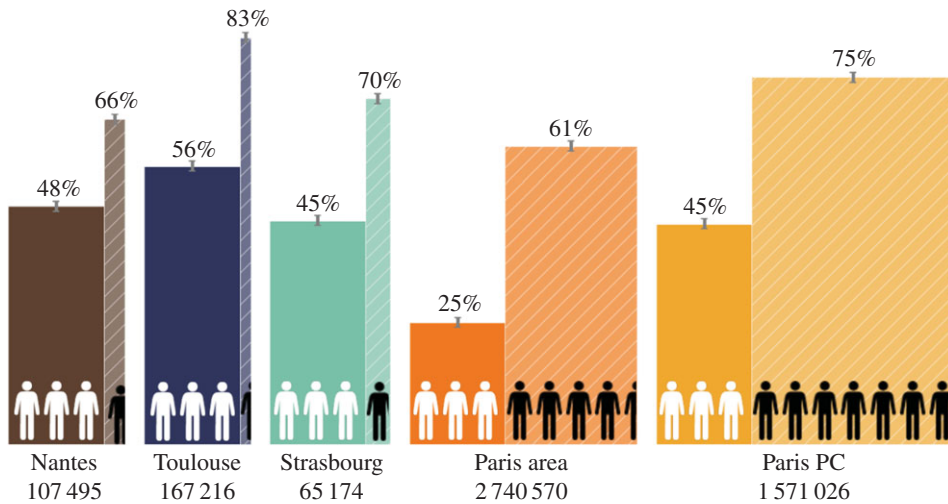
$$s_{TM} = \frac{\sum_{(a,b) \in E_C \cap E_{\text{com}}^{TM}} w_{ab}}{\sum_{(a,b) \in E_C \cup E_{\text{com}}^{TM}} w_{ab}} \quad (3.3)$$

both for  $TM = \text{car}$  and  $TM = \text{PT}$ . It represents the fraction of commuters using, respectively, car and PT, who have access to privileged PT connections (i.e. for which there exists a link corresponding to their commuting in the PT structural pattern).

The bar charts in figure 5 show the comparison between commuting flows and privileged connections for several urban agglomerations. Full bars refer to commuters choosing the car, while dashed bars refer to the choice of PT. The width of full bars are set to be equal for all the cities considered, and the width of the corresponding dashed bar is set proportionally. Hence, for each city the number of black (respectively, white) stick men over the total number of stick men corresponds to the fraction of individuals choosing to commute by PT (respectively, car). The total number of commuters (using either car or PT) in each city is indicated below each bar. For example, in the Paris central agglomeration (Paris Petite Couronne (PC), on the right of the figure), there are about 1.6 million commuters using either car or PT. Among them, for every three commuters choosing the car, about seven choose PT.

The height of the full bars (respectively, dashed filled) assigns the weighted Jaccard index  $s_{\text{car}}$  (respectively,  $s_{\text{PT}}$ ), indicating the fraction of individuals choosing the car (respectively, PT) even with access to privileged PT connections. For example, in the Paris central agglomeration, among all commuters choosing the car only 45% would have access to privileged connections, while among those who are choosing PT, 75% can rely on efficient transportation. Error bars on the top of the bars are obtained by repeating the methodology 100 times, with different random matrices initializing NMF. The small size of the error bars shows the robustness of the pattern detection.

A significant difference between the commuting practice in Paris agglomeration and other urban areas is evident. For Paris urban agglomeration, the flow of inter-municipality commuters choosing PT is larger than that of people commuting by car, in contrast to the other cities investigated. This may be partly explained by a travel time factor, which increases above the tolerated value for Toulouse, Nantes and Strasbourg (figure 3). Besides, figure 5 indicates that the fraction of commuters having access to privileged connections and actually using the PT systems is larger than the fraction of them using the car for all urban agglomerations studied. This supports our definition of privileged connections based on



**Figure 5.** Similarity between commuter flows and PT privileged connections in French municipal areas. For each of the urban agglomerations considered (Toulouse, Nantes, Strasbourg, Paris area and Paris Petite Couronne), the bar chart's height indicates the weighted Jaccard index  $s_{TM}$  between the commuter flow network  $G_{com}^{TM}$  and the PT structural pattern network  $G_C$  (for further explanation see text). The total number of commuters within each city using either car or PT is indicated below each bar.

commuting time with little variability and a limited transfer number. This corroborates the strong role of the latter factors in the decision-making to use PT or car. Furthermore, we observe that in the greater Paris area only 25% of car commuters have access to privileged transportation connections. Instead, in other cities, although more than 48% of car drivers have access to rapid connections, they still commute by car. In particular, in Toulouse a large percentage of commuters have access to good services according to the criteria introduced here, as there is large overlap between privileged connections and both PT (83%) and car (56%) commuting flows. However, there is still a non-negligible amount of people commuting by car. Based on this analysis, we can distinguish between two main trends in commuting: (i) there are cities where a large part of the population tend to do inter-municipality trips by car disregarding the quality of PT services, examples are Nantes, Toulouse and Strasbourg. (ii) On the other hand, in Paris and its agglomeration, according to the metrics introduced, there is a good agreement between the needed and provided services of public transportations. This result is supported by a pairwise comparison between the car and the PT commuting flows for every pair of municipalities (see the electronic supplementary material, section S5).

## 4. Conclusion

Efficient analysis of public transportation networks is possible via abstract representations, which in turn help us to reveal hidden characteristics of such systems. As our main scientific contribution we provided a solution for this challenge by introducing a novel description, which combines multi-edge and P-space representations of multilayer transportation networks. We characterize these systems from the user's point of view through a description, which is detached from constraints imposed by their spatial embeddedness, but which incorporates their temporal variance. To further develop our framework, we adjusted earlier-defined methods and used them to identify effective routes and hidden transportation patterns, which were not evidently built due to overall design. We found cohesive and bipartite patterns of privileged connections induced by different ways of access of far-apart urban areas in French municipals such as greater Paris, Toulouse, Nantes or Strasbourg. We further analysed the overall efficiency of the corresponding transportation systems when compared with the commuting flow. We found that while the transportation system of Paris is somewhat meeting overall demands and its use is preferred over the car alternative, in smaller cities, the transportation systems may not meet user expectations, leaving room for improvement, and even people with access to fast transportation options prefer to use car instead.

We made some assumptions during our study, which set some limitations on the generalization of our results. First of all, we considered only a 3 h time window to build our user-based representation. Extending this time window or considering different periods would potentially highlight further transportation patterns, assigning a direction to explore in the future. Furthermore, we operated with average frequencies of services neglecting the effect of any perturbation in the transportation system. This was a valid approach in our case as no major variance was observed during the analysed period. Nevertheless, to get around this limitation one can easily adjust our definition such that it considers dynamically unexpected perturbations on each line. Finally, we assumed that passengers go blindly to a stop without considering that certain passengers attempt to reduce their waiting time by timing their arrival at transit stops to an optimal period before vehicle departure. On the other hand, this can be easily considered in our representation by introducing arrival times of users, e.g. depending on the frequency of the first line they take. In addition, note that aspects such as adaptive travelling behaviour or the prediction of individual mobility patterns are out of the scope of the present methodology but they indicate possible future directions of research.

Several extensions of our methodology are possible. Parameters like the periods in focus, length of observations, number of transfers, etc., can be tailored for other systems, while a further refinement is possible by considering needs of various types of users. Our way of characterization of privileged connections may be used to profile and compare different transportation systems to disclose generalities in their design. The use of this methodology in the future could help to enhance resilience of local transportation systems to provide better design policies for future developments.

**Data accessibility.** All data used in this paper are openly shared by the local transportation companies or central statistical institutes. The datasets used can be found within an extensive list of similar datasets collected in [16,17], while the census data used can be found in [43,52]. The source of the implementation of our methodology is openly shared under the link [https://github.com/lalessan/user\\_basedPT](https://github.com/lalessan/user_basedPT).

**Authors' contributions.** All authors designed the research, participated in writing and approved the final manuscript. L.A. completed the data analysis and performed the numerical simulations.

**Competing interests.** The authors claim no financial or non-financial competing interests.

**Funding.** The research was partially supported from the Lagrange Project funded by the CRT Foundation, and the FET Multiplex Project (EU-FET-317532) funded by the European Commission. L.A. is grateful to the DANTE Inria team and the ISI Foundation for funding her internship and for ENS de Lyon for the Ampère Excellence fellowship. The funding bodies had no role in study design, data collection and analysis, preparation of the manuscript, or the decision to publish.

**Acknowledgements.** We thank the reviewers of our work for their constructive comments.

## References

- Albeverio S, Andrey D, Giordano P, Vancheri A (eds). 2008 *The dynamics of complex urban systems: an interdisciplinary approach*. Heidelberg, Germany: Physica.
- Sheffi Y. 1985 *Urban transportation networks*. Englewood Cliffs, NJ: Prentice-Hall.
- Bell MGH, Iida Y. 1997 *Transportation network analysis*, 1st edn. New York, NY: Wiley.
- Sen P, Dasgupta S, Chatterjee A, Sreeram PA, Mukherjee G, Manna SS. 2003 Small-world properties of the Indian railway network. *Phys. Rev. E* **67**, 036106. (doi:10.1103/PhysRevE.67.036106)
- Kennedy C, Miller E, Shalaby A, Maclean H, Coleman J. 2005 The four pillars of sustainable urban transportation. *Transp. Rev.* **25**, 393–414. (doi:10.1080/01441640500115835)
- Mandl CE. 1980 Evaluation and optimization of urban public transportation networks. *Eur. J. Oper. Res.* **5**, 396–404. (doi:10.1016/0377-2217(80)90126-5)
- Banavar JR, Maritan A, Rinaldo A. 1999 Size and form in efficient transportation networks. *Nature* **399**, 130–132. (doi:10.1038/20144)
- Bates J, Polak J, Jones P, Cook A. 2001 The valuation of reliability for personal travel. *Transp. Res. E-Log.* **37**, 191–229. (doi:10.1016/S0950-0804(00)00011-9)
- Carey M. 1998 Optimizing scheduled times, allowing for behavioural response. *Transp. Res. B-Methods* **32**, 329–342. (doi:10.1016/S0191-2615(97)00039-8)
- Van Oort N, Van Nes R. 2009 Regularity analysis for optimizing urban transit network design. *Publ. Transp.* **1**, 155–168. (doi:10.1007/s12469-009-0012-y)
- Lambert JH, Sarda P. 2005 Terrorism scenario identification by superposition of infrastructure networks. *J. Infrastructure Syst.* **11**, 211–220. (doi:10.1061/(ASCE)1076-0342(2005)11:4(211))
- 2016 TCRP synthesis 115. Transportation Research Board. See [http://onlinepubs.trb.org/Onlinepubs/tcrp/tcrp\\_syn\\_115.pdf](http://onlinepubs.trb.org/Onlinepubs/tcrp/tcrp_syn_115.pdf) (accessed 16 February 2016).
- Pelletier MP, Trépanier M, Morency C. 2011 Smart card data use in public transit: a literature review. *Transp. Res. Part C: Emerg. Technol.* **19**, 557–568. (doi:10.1016/j.trc.2010.12.003)
- Furth PG, Hemily BJ, Muller T, Strathman JG. 2003 Uses of archived AVL-APC data to improve transit performance and management: review and potential. Transportation Research Board, Washington, DC, USA. See [http://onlinepubs.trb.org/onlinepubs/tcrp/tcrp\\_rpt\\_113.pdf](http://onlinepubs.trb.org/onlinepubs/tcrp/tcrp_rpt_113.pdf).
- Steenbruggen J, Borzacchiello MT, Nijkamp P, Scholten H. 2013 Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. *GeoJournal* **78**, 223–243. (doi:10.1007/s10708-011-9413-y)
- Google Code. *The GoogleTransitDataFeed Open Source Software project*. See <https://code.google.com/archive/p/googletransitdatafeed/wikis/PublicFeeds.wiki> (accessed 30 June 2016).
- Google Groups. *General transit feed spec changes*. See <https://groups.google.com/forum/#!forum/gtfs-changes> (accessed 30 June 2016).
- Zhao Y. 2000 Mobile phone location determination and its impact on intelligent transportation systems. *IEEE Trans. Intell. Transp. Syst.* **1**, 55–64. (doi:10.1109/6979.869021)
- El-Geneidy AM, Horning J, Krizek KJ. 2011 Analyzing transit service reliability using detailed data from automatic vehicular locator systems. *J. Adv. Transp.* **45**, 66–79. (doi:10.1002/atr.134)
- Shen Y, Xu J, Zeng Z. 2015 Public transit planning and scheduling based on AVL data in China. *Int. Trans. Operat. Res.* **2015**, 1–23. (doi:10.1111/itor.12164)
- Mesbah M, Lin J, Currie G. 2015 'Weather' transit is reliable? Using AVL data to explore tram performance in Melbourne, Australia. *J. Traffic*

- Transp. Eng.* **2**, 125–135. (English Edition). (doi:10.1016/j.jtte.2015.03.001)
22. Mazloumi E, Currie G, Sarvi M. 2008 Assessing measures of transit travel time variability and reliability using AVL data. *Transportation Research Board Annual Meeting*, no. 87 (Washington, DC, USA). See <http://trid.trb.org/view.aspx?id=1152777>.
  23. Derrible S, Kennedy C. 2009 Network analysis of world subway systems using updated graph theory. *Transp. Res. Record: J. Transp. Res. Board* **2112**, 17–25. (doi:10.3141/2112-03)
  24. Sienkiewicz J, Hołyst JA. 2005 Statistical analysis of 22 public transport networks in Poland. *Phys. Rev. E* **72**, 046127. (doi:10.1103/PhysRevE.72.046127)
  25. Levinson D. 2012 Network structure and city size. *PLoS ONE* **7**, e29721. (doi:10.1371/journal.pone.0029721)
  26. Louf R, Roth C, Barthelemy M. 2014 Scaling in transportation networks. *PLoS ONE* **9**, e102007. (doi:10.1371/journal.pone.0102007)
  27. Louf R, Barthelemy M. 2014 How congestion shapes cities: from mobility patterns to scaling. *Sci. Rep.* **4**, 5561. (doi:10.1038/srep05561)
  28. Kivela M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA. 2014 Multilayer networks. *J. Complex Netw.* **2**, 203–271. (doi:10.1093/comnet/cnu016)
  29. Boccaletti S, Bianconi G, Criado R, del Genio CI, Gómez-Gardeñes J, Romance M, Sendiña-Nadal I, Wang Z, Zanin M. 2014 The structure and dynamics of multilayer networks. *Phys. Rep.* **544**, 1–122. (doi:10.1016/j.physrep.2014.07.001)
  30. Tsiotas D, Polyzos S. 2015 Decomposing multilayer transportation networks using complex network analysis: a case study for the Greek aviation network. *J. Complex Netw.* **3**, 642–670. (doi:10.1093/comnet/cnv003)
  31. Gallotti R, Barthelemy M. 2015 The multilayer temporal network of public transport in Great Britain. *Sci. Data* **2**, 140056. (doi:10.1038/sdata.2014.56)
  32. Holme P, Saramäki J. 2012 Temporal networks. *Phys. Rep.* **519**, 97–125. (doi:10.1016/j.physrep.2012.03.001)
  33. Cormen TH, Leiserson CE, Rivest RL, Stein C 2001 *Introduction to algorithms*, 2nd edn. Cambridge, MA: The MIT Press.
  34. Asperges T *et al.* 2007 Déterminants des choix modaux dans les chaînes de déplacements. [Determinants of transport mode choice.] Résumé, Plan d'Appui scientifique à une politique de Développement Durable (PADD II), Partie 1. [Summary, Plan of scientific support to a sustainable development policy]. See <http://goo.gl/qF75Jx>.
  35. Le Jeanic T *et al.* 2010 La mobilité des Français, panorama issu de l'enquête nationale transports et déplacements 2008 [Mobility in France, an overview from the national survey of transport and mobility 2008]. Paris: ministère de l'Écologie, du Développement durable, des Transports et du Logement. [The Ministry for Ecology, Sustainable Development, Transport and Housing]. See <http://goo.gl/JWp10H>.
  36. Lillo Viedma FE. 2011 Coloured-edge graph approach for the modelling of multimodal networks. PhD thesis, Auckland University of Technology.
  37. Sen P, Dasgupta S, Chatterjee A, Sreeram PA, Mukherjee G, Manna SS. 2003 Small-world properties of the Indian railway network. *Phys. Rev. E* **67**, 036106. (doi:10.1103/PhysRevE.67.036106)
  38. Sienkiewicz J, Hołyst JA. 2005 Statistical analysis of 22 public transport networks in Poland. *Phys. Rev. E* **72**, 046127. (doi:10.1103/PhysRevE.72.046127)
  39. Von Ferber C, Holovatch T, Holovatch Y, Palchykov V. 2009 Public transport networks: empirical analysis and modeling. *Eur. Phys. J. B* **68**, 261–275. (doi:10.1140/epjb/e2009-00090-x)
  40. Jolliffe JK, Hutchinson TP. 1975 A behavioural explanation of the association between bus and passenger arrivals at a bus stop. *Transp. Sci.* **9**, 248–282. (doi:10.1287/trsc.9.3.248)
  41. Csikos D, Currie G. 2007 Investigating consistency in passenger arrivals: insights from longitudinal ticket validations. In *Conf. of Australian Institute of Transport Research (CAITR)*, 29th, 2007, Adelaide, South Australia, Australia. See <http://trid.trb.org/view.aspx?id=868843>.
  42. Chang S, Hsu CL. 2001 Modeling passenger waiting time for intermodal transit stations. *Transp. Res. Record: J. Transp. Res. Board* **1753**, 69–75. (doi:10.3141/1753-09)
  43. National Institute of Statistics and Economic Studies, France (INSEE). 2008 *Opération statistique: Enquête Nationale Transports et Déplacements 2007–2008*. See <http://goo.gl/5CZ6r1> (accessed 30 June 2016).
  44. Psorakis I, Roberts S, Ebdon M, Sheldon B. 2011 Overlapping community detection using Bayesian non-negative matrix factorization. *Phys. Rev. E* **83**, 066114. (doi:10.1103/PhysRevE.83.066114)
  45. Wang F, Li T, Wang X, Zhu S, Ding C. 2011 Community discovery using nonnegative matrix factorization. *Data Min. Knowl. Disc.* **22**, 493–521. (doi:10.1007/s10618-010-0181-y)
  46. Zhang S, Wang RS, Zhang XS. 2007 Uncovering fuzzy community structure in complex networks. *Phys. Rev. E* **76**, 046103. (doi:10.1103/PhysRevE.76.046103)
  47. Psorakis I, Roberts S, Sheldon B. 2010 Efficient Bayesian community detection using non-negative matrix factorisation. See <http://arxiv.org/abs/1009.2646>.
  48. Zhang ZY, Wang Y, Ahn YY. 2013 Overlapping community detection in complex networks using symmetric binary matrix factorization. *Phys. Rev. E* **87**, 062803. (doi:10.1103/PhysRevE.87.062803)
  49. He D, Jin D, Baquero C, Liu D. 2014 Link community detection using generative model and non-negative matrix factorization. *PLoS ONE* **9**, e86899. (doi:10.1371/journal.pone.0086899)
  50. Gao X, Wang X, Jin D, Cao Y, He D. 2014 The (un)supervised detection of overlapping communities as well as hubs and outliers via (bayesian) NMF. In *Proc. of the Companion Publication of the 23rd Int. Conf. on World Wide Web, WWW '14 Companion*, pp. 233–234. New York, NY: ACM. (doi:10.1145/2567948.2577307)
  51. Owen AB, Perry PO. 2009 Bi-cross-validation of the SVD and the non-negative matrix factorization. *Ann. App. Stat.* **3**, 564–594. (doi:10.1214/08-AOAS.227)
  52. National Institute of Statistics and Economic Studies, France (INSEE). 2010 *Fichier Mobilités professionnelles des individus: déplacements commune de résidence / commune de travail*. [Home to work commuting flows by transportation modes]. See <http://goo.gl/jyJqF8>.