TRANSPORT
ISSN 1648-4142 / eISSN 1648-3480
2020 Volume 35 Issue 6: 576–587
https://doi.org/10.3846/transport.2020.14159

SPATIOTEMPORAL DYNAMICS OF PUBLIC TRANSPORT DEMAND: A CASE STUDY OF RIGA

Dmitry PAVLYUK*, Nadežda SPIRIDOVSKA, Irina YATSKIV (JACKIVA)#

Dept of Mathematical Methods and Modelling, Faculty of Engineering, Transport and Telecommunication Institute, Riga, Latvia

Received 6 July 2020; revised 17 August 2020; accepted 2 September 2020

Abstract. Sustainable urban mobility remains an emerging research topic during last decades. In recent years, the smart card data collection systems have become widespread and many studies have been focused on usage of anonymized data from these systems for better understanding of mobility patterns of Public Transport (PT) passengers. Data-driven mobility patterns can benefit transport planners at strategic, tactical, and operational levels. A particular point of interest is a spatiotemporal dynamics of mobility patterns that highlights transformation of the PT passenger flows over the time continuously or in response to modifications of the PT system and policies. This study is aimed to estimation and analysis of the spatiotemporal dynamics of PT passenger flows in Riga (Latvia). A multi-stage methodology was proposed and includes three main stages: (1) estimation of individual trip vectors, (2) clustering of trip vectors into spatiotemporal mobility patterns, and (3) further analysis of mobility patterns' dynamics. The best practice methods are applied at every stage of the proposed methodology: the smart card validation flow is used for extracting information on boarding locations; the trip chain approach is used for estimation of individual trip destinations; vector-based clustering algorithms are utilised for identification of mobility patterns and discovering their dynamics. The resulting methodology provides an advanced tool for observing and managing of PT demand fluctuation on a daily basis. The methodology was applied for mining of a large smart card data set (124 million records) for year 2018. Most important empirical results include obtained daily mobility patterns in Riga, their clusters, and within-cluster dynamics over the year. Obtained daily mobility patterns allows estimation of a city-level PT origin–destination matrix that is useful in many applied areas, e.g., dynamic passenger flow assignment models. Mobility pattern-based clustering of days allows effective comparison and flexible tuning of the PT system for different days of a week, public holidays, extreme weather conditions, and large events. Dynamics of mobility patterns allows estimating the effect of implementing changes (e.g., fare increase or road maintenance) and demand forecasting for user-focused development of PT system.

Keywords: user travel behaviour, transport modelling, big data, public transport, smart card data, clustering.

Introduction

The key goal of the Sustainable Urban Mobility Plan (SUMP) (EC 2013) is improving the accessibility of urban areas and providing high-quality and sustainable mobility. SUMP is defined (Rupprecht *et al.* 2019) as a strategic plan designed to satisfy the mobility needs of people and businesses in cities and their surroundings for a better quality of life. Sustainable urban development requires an appropriate transport system, so mobility is an integral part of the development of the entire city. Urban mobility is becoming a critical challenge with the increasing population in cities and the growing urbanisation level. Seeing these trends, attention to mobility alternatives for meeting the demand of growing populations should be considered in the perspective of sustainable transport alternatives.

Smart card systems, adopted for Public Transport (PT) in many cities over the world, intensively collect a large volume of data that have a huge potential for revealing the mobility demand. A lot of efforts have been made in scientific literature to develop tools for extracting and utilizing information from smart card databases. Smart card data can be used for strategic, tactical and operational tasks and researchers agree that more can be done with the data: destination estimation, network performance, travel behaviour. A regular smart card database stores information at the lowest level of disaggregation (usually, individual smart card validations), so there is a long methodological way to aggregating them into mobility patterns. A typical methodology includes many intermediate steps (like

*Corresponding author. E-mail: *dmitry.pavlyuk@tsi.lv*

#Editor of the TRANSPORT – the manuscript was handled by one of the Associate Editors, who made all decisions related to the manuscript (including the choice of referees and the ultimate decision on the revision and publishing).

estimation of boarding and alighting stops, identification of trip chains, etc.), which are associated with many hypotheses made and technical problems raised. Frequently, these hypotheses and problems are city- and data-specific, so case studies play an important role in PT researches.

This paper aims to define the mobility patterns of the users of PT in Riga (Latvia) based on smart card data. The Riga smart card system provides only entry validations and is not integrated with scheduling databases and PT vehicle locations. This fact challenges the utilised methods, as well as the adopted hypotheses that need to be verified in different contexts. Methodological and technical challenges that appear in consecutive stages of mobility pattern recognition were discussed and general feasibility of the methodology for solving strategic-level problems was demonstrated. The paper is organised as follows: Section 1 presents a brief overview of spatiotemporal PT mobility studies; Section 2 introduces details of our methodology; Section 3 describes the case study, obtained empirical results and related discussion; Section 4 provides a discussion, and the last section summarizes the conclusions.

1. State of the art

Better understanding of mobility behaviours is highly required for services customization and more effective PT systems. Many original studies have been published over the last two decades: Pelletier *et al.* (2011), Faroqi *et al.* (2018), and Welch, Widita (2019) composed extensive literature reviews that cover different aspects of PT smart card data processing. Although many authors widely refer their studies as spatiotemporal analysis of urban mobility patterns, the essence of this analysis could be different.

The first step towards spatiotemporal mobility pattern analysis is based on discovering the spatial distribution of PT boardings and its dynamics over the research period. Morency *et al.* (2007) analysed the number of boardings at the PT stop level during 277 days within a Canadian transit network; Zhong *et al.* (2015) utilised temporal patterns of boardings during the days and implemented correlation analysis of PT stops for one-week Singapore smart card data; Briand *et al.* (2017) used the temporal profile of boardings for clustering passengers and discover their behaviour; El Mahrsi *et al.* (2017) implemented a similar methodology to discover patterns of different smart card types.

Although boarding activities are important for transit operators, they do not reveal mobility directions and trajectories. Therefore, many researchers put their efforts on estimation of individual trip destinations. The problem is usually complicated by the absence of data on the alighting PT stop – many smart card usage policies require validation of the card at the boarding only. In addition to estimation of alighting stops, the problem of transfer stops arises: a data processing algorithm should distinguish final destinations of passengers from intermediate ones. Typically, this problem is solved on the base of the next smart card validation and predefined rules (e.g., a time period threshold between consecutive validations). These rules are the subject of selection and can be quite complicated – for example, a short stop for carrying a child to school and continuing the way to the office could be considered as a destination of interest for mobility patterns. The trip chain approach is widely used in many studies to extend spatial boarding information (Gentile, Noekel 2016). Trépanier *et al.* (2007) provided an algorithm for trip destination estimation and applied it to the analysis of spatiotemporal patterns of route loadings. Wang *et al.* (2011) estimated origin and destination of passengers for evaluation of bus connections. Tao *et al.* (2014) compared origins and destinations distribution over the time to discover patterns for different social groups of passengers. Similar to studies on boarding information, origin and destination pairs were recently used for passenger and weeks clustering (Deschaintres *et al.* 2019; He *et al.* 2020).

Many researches are focused on estimation of PT passengers' behaviour on individual, boarding and alighting stop, and route levels, and only a very limited number of studies are devoted to analysis of a large-scale variability of PT demand using smart card data. Ma *et al.* (2013) analysed 5-day smart card data in Beijing (China), for discovering the temporal and spatial regularity of PT trips. The *k*-means algorithm was used to identify clusters of passengers with a similar level of demand regularity. Similarly to researches by Ma *et al.* (2013) and Kieu *et al.* (2015) identified temporal patterns of PT demand using 4-month smart card data for South East Queensland (Australia) and applying the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering. Briand *et al.* (2017) investigated long-term (year-to-year) changes of PT travel behaviour for 5 years of smart card data from Gatineau (Canada). They applied a Gaussian mixture model for cluster identification and discovered a relative stability of cluster characteristics over years. Goulet-Langlois *et al.* (2016) analysed 4-week PT passengers' activities in London (UK) extracted several user clusters and discovered associations between travel patterns and socio-demographic attributes. Later in research by Goulet-Langlois *et al.* (2018), the authors used entropy rates for performing the analysis of cluster regularities and variability over time. Manley *et al.* (2018) also performed a research of London smart card data for discovering spatiotemporal variation of travel demand. Similarly, to previous studies, the analysis was conducted on an individual level: the authors clustered passengers using similarities of their PT usage. Several recent studies were focused on combination of smart card data and other sources of travel information for discovering mobility patterns and their variability. Long and Thill (2015) combined smart card data with data from the household travel survey in Beijing for discovering typical commuting trips and their variability over time. Qi *et al.* (2019) enhanced smart card data with Points Of Interests (POI) for constructing patterns of regional mobility. Chen *et al.* (2020) also examined regional mobility patterns by enhancing smart card data with and mobile internet data. Recently, Egu and Bonnel

(2020) analysed day-to-day variability of mobility patterns using 6-month smart card data from Lyon (France). They proposed an intrapersonal variability metric that allows better understanding of spatiotemporal patterns in smart card data.

One of the best practices in the spatiotemporal urban mobility analysis is managing discovered individual origin–destination (OD) trips as vectors. These vectors are usually grouped on the base of Traffic Analysis Zones (TAZ) and further estimation of OD matrixes. This approach is widely used in recent studies by Barry *et al.* (2009), Kumar *et al.* (2018), Munizaga, Palma (2012), Wei *et al.* (2017), due to the simplicity of the resulting structure and wide range of visualisation tools. At the same time, it requires a strong assumption on the correct TAZ definition. Recently, it was argued that the preliminary TAZ definition may be unreasonable and an assumption-free approach is highly required (Xu *et al.* 2019). The authors suggested to use clustering of POI for mining the zones from data and applied this data-driven zone definition for spatiotemporal analysis of bicycle riders' mobility pattern.

2. Research methodology

The methodology, proposed in this study, covers different levels of smart card data analysis – from raw validation data to dynamics of spatiotemporal data-driven mobility patterns.

2.1. Data structure

The key input for the research methodology is a structure of available data and its completeness. In this study, data from two non-integrated data sources were utilized:

- smart card validation data set that is provided by PT operator;
- scheduling information that was collected in General Transit Feed Specification (GTFS) format from *Open Mobility Data* (OMD 2019).

A sample of validation records, obtained from the smart card system contains the following information:

- *TransactionDatetime* contains an accurate card validation timestamp;
- *RouteID* codes information about the transport mode and the route number (for example, *tram*_11);
- *RunDirection* is a binary indicator of movement direction (forward or backward);
- *CardID* is a unique card identifier that is utilised for trip chain identification;
- *CardType* includes information about applied fares (one-hour tickets, discount tickets, etc.);
- *TransportID* is a unique PT vehicle number.

The smart card database is not integrated with the automated vehicle location system, so there is no natural information about the spatial dimension of data. Thus, PT schedules were utilised and data fusion was implemented for enhancing the smart card data set with spatial information. Scheduling information for every date is collected via historical feeds in GTFS format and includes information on:

- *RouteID* and *TripDirection* that are matched to corresponding fields from the smart card database;
- *StopSpatialLocation* includes geographical coordinates of every PT stop in the system;
- *ScheduledTripID* is a unique identifier of a scheduled PT trip;
- *StopTimes* are scheduled arrival and departure times for every trip and PT stop.

The main stages of the research methodology are:

- trip vector construction:
  - fusion of data from smart card and scheduling databases;
  - estimation of a boarding stop for every individual trip leg;
  - trip chain recognition and trip vector construction;
- mobility pattern estimation:
  - discovering typical mobility patters by spatiotemporal clustering of trip vectors;
- analysis of mobility patterns' dynamics over time:
  - clustering sample dates by mobility patterns' similarity;
  - estimation of within-cluster variability of mobility patterns during the research period (a year).

Many existing studies address a specific stage of smart card data analysis – spatial analysis of boardings, estimation of alighting stop and construction and trip vector, PT demand variation, etc. Every stage of smart card data analysis requires specific assumptions and is associated with specific errors and information losses. Our research methodology is a composite one, so it naturally aggregates assumptions and errors of intermediate stages. Thus, the case study in addition to its practical added value, allows revealing feasibility of complex smart card data analysis for strategic level problem solving and decision making.

2.2. Trip vector construction

2.2.1. Data fusion

Due to disintegrated smart card validation and scheduling databases, the first step was discovering of actual PT run from the smart card database and matching them to scheduled PT runs (run is defined as the movement of the PT vehicle between end stops of a route). The smart card database includes information on transport vehicle numbers and moving directions, but does not contain identifiers of the PT runs. To discover the actual PT run, the following rules were proposed:

- a new PT run is started if the direction is changed (from forward to backward and reverse);
- a time lag between two consecutive registrations exceeds 30 min (this works for specific cases where a transport vehicle goes to a park for maintenance and return to the same starting point);
- PT runs with one smart card validation are excluded as erroneous.

When actual PT runs are extracted, they are represented by a tuple *ObservedRun*:

$$ObservedRun_r = \{RouteNumber_r, RunDirection_r, FirstRegistration_r, LastRegistration_r\}, \quad (1)$$

where: r is a PT run index; *RouteNumber*, *RunDirection* correspond to the extracted run; *FirstRegistration*, *LastRegistration* are calculated as timestamps of first and last smart card validation during the run.

Next, scheduled runs *ScheduledRun* are extracted from the scheduling database as:

$$ScheduledRun_{rs} = \{RouteNumber_{rs}, RunDirection_{rs}, FirstStop_{rs}, LastStop_{rs}\}, \quad (2)$$

where: rs is a scheduled run index; *FirstStop*, *LastStop* are scheduled departure timestamps for the first and last stops in the run.

Finally, the matching rule for observed and scheduled PT runs is introduced as:

$$\begin{aligned} &RouteNumber_r = RouteNumber_{rs} \text{ AND} \\ &RunDirection_r = RunDirection_{rs} \text{ AND} \\ &FirstStop_{rs} > FirstRegistration_r - Sh1 \text{ AND} \\ &LastStop_{rs} > LastRegistration_r + Sh2 \text{ AND} \\ &\min_{rs} w\left|FirstRegistration_r - FirstStop_{rs}\right| + \\ &(1-w)\cdot\left|LastRegistration_r - LastStop_{rs}\right|, \end{aligned} \quad (3)$$

where: $Sh1$, $Sh2$ are tolerance thresholds for run start and end timestamps; w is a relative confidence for first and last stop matching.

The nature of $Sh1$ and $Sh2$ tolerance thresholds is different: $Sh1$ corresponds to a time lag between PT transport departure from the beginning stop of the route and the first smart card validation, while $Sh2$ corresponds to a time lag between the last validation and the end stop of the route. The first time lag is usually shorter – in many cases, the beginning stop is used by many passengers, so the first validation happens immediately after the transport departure. The second time lag is usually longer – boardings during last stops of the route are rare. Thus, $Sh1$ is arbitrary set to 10 min and $Sh2$ set to 60 min in this case study. In addition to the tolerance thresholds, which are designed for cutting out incorrect runs, the parameter w was introduced and allows balancing the importance of first and last stop matching. In this case study, w value was to 0.5 (equal importance of the first and the last stop); other tested values do not affect final results significantly. Potentially, this value can be important for solving problems, where precise recognition of the PT run is required. In our case, an improper selection of a PT run from a series of consecutive PT runs, serving the same route with a short time interval (2…3 min) is not critical for estimation of longer trips.

The suggested rule – Equation (3) – for PT run selection does not control the uniqueness of runs, so a scheduled run may be matched to several observed ones (for example, in the case of buses, gathered together after a traffic jam). This approach is functional until the researchers are interested in actual times of PT stops only.

2.2.2. Boarding stop estimation

The problem of boarding stop estimation is recently addressed in several studies (Barry *et al.* 2009; Chen, Fan, 2018; Wang *et al.* 2011). Following Chen and Fan (2018), schedule information was matched with actual card validation timestamps to associate a trip with a boarding stop. Scheduled and actual *StopTimes* can be different due to unexpected changes in PT operation (e.g., delays due to congested traffic conditions). To overcome this issue, scheduled stop times was corrected utilizing the information about the smart card validation flow (assuming that validations are temporary clustered in a short time after the actual stop timestamp).

2.2.3. Trip chain recognition and trip vector construction

The problem of alighting stop estimation is a complex one and widely acknowledged in literature. In this study, a popular trip chain methodology was applied for identification of intermediate and final destinations. Following the methodology, we grouped consecutive PT legs into a single trip. Within the scope of this research, the primary goal is defined as identification of mobility vectors, thus actual alighting stop information is not absolutely necessary. Instead of estimation of PT trip's alighting stop, the next boarding stop was used for trip vector construction (Figure 1).

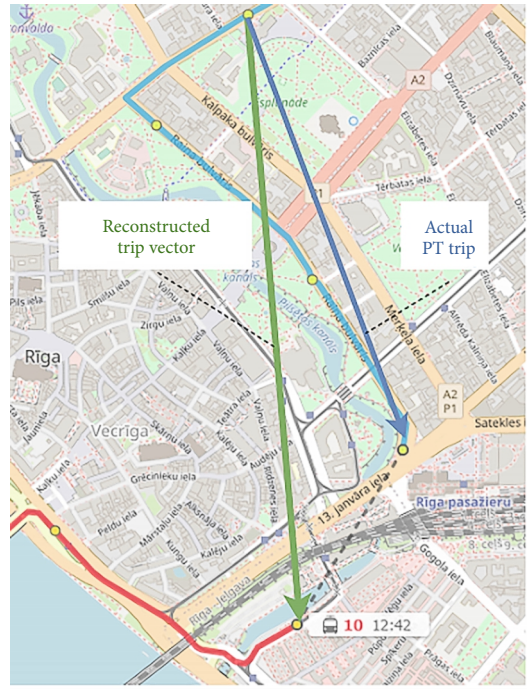

Figure 1. Illustration of a difference between actual and reconstructed PT trip vectors (the actual vector is presented in blue, the reconstructed one – in green)

Our analysis was focused on chained trips only, so the boarding stop of the following leg was arbitrary used as a destination for the previous one. Thus, the step of alighting stop identification, which is usually based on the nearest distance between a stop from the previous leg and the boarding stop of the following one, was skipped. In addition, at this stage all smart card validations that happen only once a day (including one-travel tickets) were excluded. The trip chain analysis is usually focused on "home – activity (work) – home" commuters, so a destination of the last leg within the day was defined as the boarding stop of the first trip. As this assumption could be weak and not supported by previous studies for Riga, these "return" trips were separately analysed.

Finally, trip vectors were defined by grouping consecutive legs. The legs are considered as consecutive if a difference between their boarding times does not exceed 60 min (this value is arbitrary selected by the Riga PT operator, which provides one-hour tickets with unlimited number of transfers). The resulting dataset includes geographical information about trips, associated with the origin (boarding stop) and destination (next leg's boarding time) and the boarding timestamp, represented as a vector *TripVector*:

$$TripVector_i = \{BoardingTime_i, OriginStop_i, DestinationStop_i\}, \quad (4)$$

where: i is a trip vector index.

2.3. Mobility pattern estimation

The dataset of trip vectors contains information about a large number of individual trips, which are usually grouped for easier visualisation and mobility pattern recognition. At this stage, the clustering technique was applied for grouping individual trip vectors into mobility vectors. The mobility vectors *MobilityVector* represent main spatiotemporal patterns of PT usage and includes information about average origin and destination locations, associated boarding timestamp, and flow volume:

$$MobilityVector_j = \{BoardingTime_j, Origin_j, Destination_j, Flow_j\}, \quad (5)$$

where: j is a mobility vector index; *Origin*, *Destination* are the geographical coordinates (longitude and latitude) of mobility vector's initial and terminal points.

Figure 2 represents an example of constructed mobility patterns for one PT route at different time periods of the day.

Note that unlike origin and destination of the trip vector, which are associated with specific PT stops, origin and destination of the mobility vector are arbitrary geographical coordinates that are not directly linked to PT stops and routes.

This approach to mobility vector construction differs our methodology from many existing studies. Many authors represent a daily mobility pattern via distribution of passengers over the spatial (stop locations) and temporal (hours) dimensions (El Mahrsi *et al.* 2017; Morency *et al.* 2007; Xu *et al.* 2019); others identify patterns via flows between predefined TAZ (Kumar *et al.* 2018). At the same time, the first approach does not utilize information on mobility directions, while the second one requires aggregation at the TAZ level. Note that the trip vectors were clustered, which differs this research from previous studies, focused on clustering of passengers (Han, Sohn 2016; Morency *et al.* 2007) or stops (Cats *et al.* 2015). The suggested approach allows constructing mobility patterns that represent spatiotemporal information about PT demand in a flexible manner: the resulting patterns may vary over the time and not linked to predefined city zones.

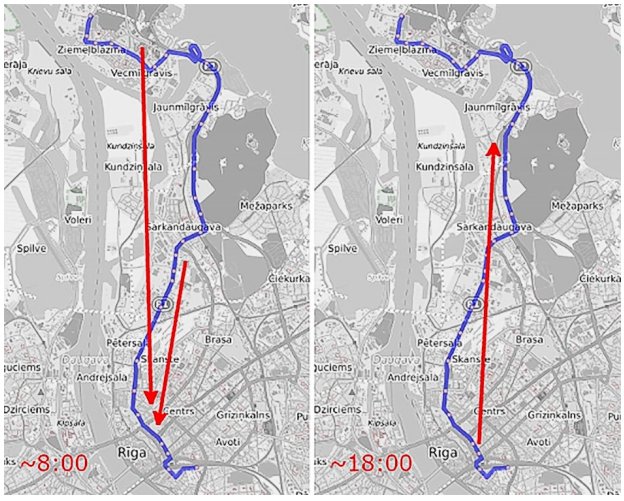

Figure 2. Typical mobility patterns for a selected bus route

Any clustering algorithms can be applied for trip vector clustering, e.g., k-means clustering with a predefined number of clusters; agglomerative hierarchical clustering with posterior identification of clusters, or DBSCAN algorithm. The clustering is executed on a daily basis, so a daily mobility pattern was constructed as a set of mobility vectors (clustered trip vectors) for every date in the research sample.

2.4. Dynamics of estimated mobility patterns

Although sample-wise smart card data provides useful information about regular travel patterns, day-to-day estimation of mobility patterns allows better understanding of pattern variability over time (Alsger *et al.* 2018). This is highly expected that mobility patterns differ for weekdays and weekends, as well as for different seasons, weather conditions, etc. Thus, the most interesting aspect of PT demand variability is mobility pattern changes within groups of similar dates. For example, from practical point of view, it can be interesting to identify changes of mobility patterns for summer weekends with good weather conditions – for improving PT service at these specific dates (for PT operators) or for monitoring inhabitant behaviour and preferences as a result of local policies and investments (for municipal governments). This grouping can be hard-

coded (e.g., split working days and public holidays), but, taking into account a large number of factors that affect PT demand, it would be beneficial to apply a data-driven approach to the problem. In this study, a clustering technique was applied for identification of dates with similar daily mobility patterns and further analysed within-cluster variability over the research period.

2.4.1. Clustering sample dates by mobility patterns

The daily mobility pattern is represented by a set of vectors, and, as any clustering technique requires specification of a distance metric, a custom similarity metric for two vector sets was introduced. A distance between two mobility vectors can be naturally estimated using any technique from a wide range of vector similarity metrics: Euclidean distance, cosine similarity, etc. The problem is related to absence of links between mobility vectors in patterns for two days: there is no prior information about the matching of mobility vector pairs between two dates. This matching is required for measuring the total similarity between mobility patterns of two dates. To solve this problem, application of the minimum-cost bipartite matching algorithm was proposed:

- mobility vectors of two days are matched to minimize the sum of pairwise similarities;
- resulting cost of matching is used as a distance metric between daily mobility patterns.

The minimum-cost bipartite matching is a classical combinatorial problem that can be solved by any well-known algorithm. In this study, the classical Hungarian method was applied. Results of mobility vector matching are illustrated in Figure 3.

Having the distance metric for daily mobility patterns, a matrix of distances can be constructed and the sample dates can be clustered to groups of similar mobility patterns for further analysis.

2.4.2. Estimation of within-cluster variability of mobility patterns

One of the key issues of city transport policy planning is variation of mobility patterns over the time (Alsger *et al.* 2018). PT authorities evaluate and adjust their current services following the changing demand. Usually a PT operator supports several schedules that applied for days with different mobility patterns. Thus, within-cluster variability of daily mobility patterns plays an important role and needs to be estimated. Mobility pattern variability is estimated by applying temporal moving average values of within-cluster distances over the sample to monitor and analyse trends and temporal variation of mobility patterns.

The research methodology is summarised in Figure 4.

There are several distinguishing features of the presented methodology:

- due to the disaggregated smart card and scheduling databases and absence of information on actual PT locations, the methodology includes the custom data fusion stage;
- the methodology considers daily mobility patterns as sets of non-fixed vectors. The majority of existing PT studies deal with mobility patterns from one of prespecified perspectives: spatial (distributions of boarding or alighting), temporal (daily time series of trips), or fixed vectoral (TAZ-based flows). The proposed definition of daily mobility patterns is more flexible, requires less assumptions and could better reflect actual PT demand;
- the methodology suggests the data-driven approach to discovering similarities between daily mobility patterns (for further clustering). The approach is based on the minimum-cost bipartite matching algorithm and, to the best of our knowledge, was not previously applied to urban mobility patterns' analysis;

Mobility pattern 1 March 2018

Mobility vector	Start time	Trip count	Origin (longitude)	Origin (latitude)	Destination (longitude)	Destination (latitude)
1	12:45	10407	24.1165	56.9471	24.1154	56.9616
2	09:02	10361	24.1556	56.9431	24.1603	56.9485
3	07:37	13274	24.1492	56.9610	24.1143	56.9539
4	09:32	6641	24.794	56.9576	24.0727	56.9477

Minimum cost matching

Mobility pattern 1 March 2018

Mobility vector	Start time	Trip count	Origin (longitude)	Origin (latitude)	Destination (longitude)	Destination (latitude)
1	10:15	8299	24.1487	56.9475	24.1123	56.9497
2	08:07	10412	24.0712	56.9530	24.0933	56.9550
3	11:34	9786	24.2039	56.9751	24.1482	56.9636
4	09:25	10422	24.1648	56.9416	24.1207	56.9375

Figure 3. Illustration of minimum-cost bipartite matching results for mobility patterns of two dates

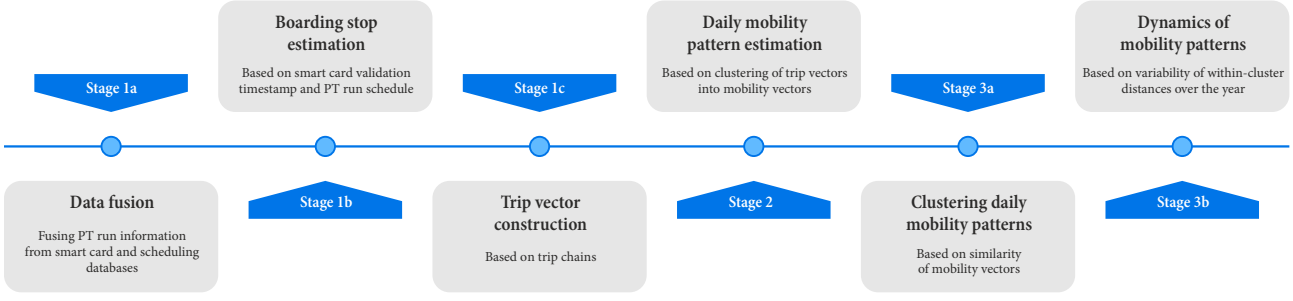

Figure 4. Stages of the research methodology

– the methodology covers the complete cycle of smart card data processing – from raw validation database to daily mobility patterns and their variability. Although best practice methods are applied at every stage, all of them are based on background assumptions and are subject to errors. The error is naturally accumulated during several consecutive stages, so the utility of the combination of best practices should be proved.

3. Case study

This study utilizes data from the Riga urban PT operator *Rīgas Satiksme* (*https://www.rigassatiksme.lv*) and its subsidiary *Rīgas Karte Ltd.* (*https://rigaskarte.lv*) that manages smart card operations. The PT in Riga runs a distance of about 45 million km and carries nearly 150 million passengers per year. Smart cards (called e-tickets) are universal electronic tickets that manage payments for PT services. They were fully introduced into the Riga PT in 2009 and they are valid for all PT modes, except suburban trains. Electronic validators are located in PT vehicles and register passengers paying for the trip. Passengers entitled to fare discounts also have to register their trips. Each validation record contains temporal (date and time), transport (PT route and direction), and fare information, which is collected at the boarding timestamp only.

A topology of the Riga PT system and distribution of served passenger flows by transport mode are presented in Figure 5. The research data set includes smart card validation records from 2018, except special public holidays when PT trips are not charged (e.g., New Year's Day, Midsummer) and several days where information is missed due to technical reasons. Thus, the data set covers 315 days with complete information. The overview of the research area and the data set is presented in Table 1.

Table 1. Description of the research area and data set

Settings	Value
Research area	Riga (Latvia)
Research area population	615 thousand inhabitants
Number of PT routes	84, including 55 bus, 19 trolleybus and 10 tram routes
Number of PT stops	1675
Time frame	1 January – 31 December 2018
Number of passengers	~124 million

The primary source of data is the smart card validation database, obtained from the Riga urban PT operator. The structure of the database is common and similar to data of other cities (Welch, Widita 2019). An example of validation records is presented in Table 2.

Note, that the smart card database contains only temporal information about validation. Although all PT transports are equipped with Global Positioning System (GPS) sensors, the smart card validation database is completely separate and does not contain any geographical (spatial) information. In addition, there is no information about PT runs (start and end times, stop times) in the database, which creates additional problems for enhancing data with spatial information. It should be mentioned that the database is not consistent enough and contains data errors, which should be identified and resolved. One of the popular errors is an absence of direction switching within a PT run, e.g., data indicate that a transport goes "forward" direction for several hours, which is impossible for the Riga PT system. These errors are ones of the background reasons for introducing a custom PT run matching rules – Equation (3).

The first stage of the methodology includes a fusion of smart card validation data within historical scheduling in-

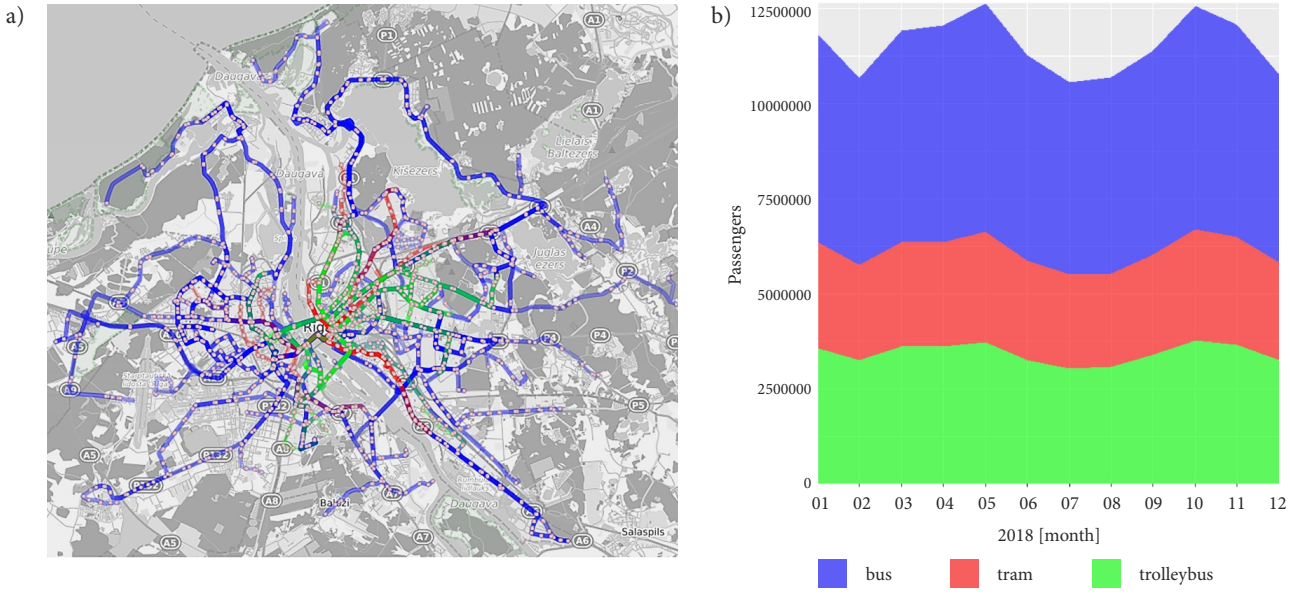

Figure 5. PT system of Riga (bus routes are coded as blue, tram routes as red, and trolleybus routes as green): a – topology of the PT of Riga; b – monthly dynamics of number of served passengers in 2018

formation, estimation of boarding stops and construction of individual trip vectors. Following the best practices, presented in Section 2, a set of trip vectors was obtained for every date in the sample. Summary statistics for intermediate steps of this stage are presented in Table 3.

The stage results were validated against smart card records of two passengers with known routes (734 validations at total) and demonstrated completeness of 67%. Missed trip vectors are mainly related to inconsistent information between smart card and scheduling databases (e.g., actual PT runs that cannot be confidently matched to scheduled). Smart cards with only one validation for a given date were excluded from consideration. Due to these reasons, the trip vectors to smart card validations ratio resulted in 42…43% (the last column in Table 3) and uniformly distributed by seasons, weekdays and day times.

Trip vectors, obtained at the first stage, are clustered into daily mobility patterns. Figure 6 presents an example set of trip vectors for a smart card and a typical daily mobility pattern.

Discovered daily mobility patterns are represented by a set of vectors of origin, destination (marked by red circles in Figure 6), average time of the day (labels), and flow volume (width of lines). Daily mobility patterns provide an extensive information on PT demand and can be used for construction and calibration of OD matrixes and making tactical decisions of PT scheduling. Different approaches to clustering were tested, including different distance met-

Table 2. Example validation records

Transaction date, time	Route ID	Run direction	Card ID	Card type	Transport ID
2 January 2018, 05:27:08	*bus*_3	forward	2030603xxx	10 one-hour tickets	544189912
2 January 2018, 05:28:07	*tram*_1	forward	6400065xxx	daily	386483920
2 January 2018, 05:21:11	*tram*_1	forward	2565066xxx	monthly	983736630

Table 3. Summary statistics of trip vector construction steps

Weekday	Smart card validations (daily average)	Validations with matched PT runs	Trip vectors	Trip vectors to validations ratio
Monday	417679	323711	179900	0.4307
Tuesday	449695	348914	196328	0.4366
Wednesday	461946	358751	202423	0.4382
Thursday	466666	362378	204147	0.4375
Friday	451185	348514	193041	0.4279
Saturday	263355	207657	114021	0.4330
Sunday	209492	164457	88966	0.4247

a) b)

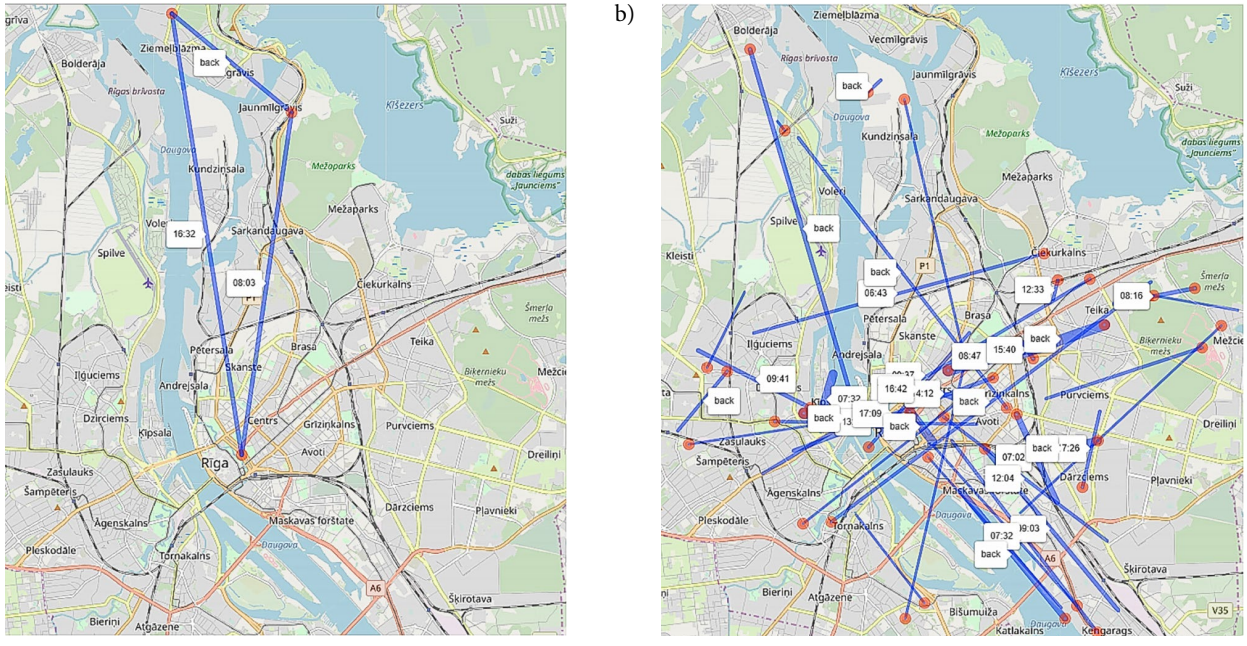

Figure 6. Typical individual trip vectors (a) and a daily mobility pattern (b) – destinations are marked with red circles; flow volumes presented by vector widths

rics and algorithms. The results are fairly similar, so for brevity reasons the following discussion is presented for one solution only – Euclidean distance-based *k*-means algorithms, executed separately for morning and evening rush hours, midday, and "return" trips (where exact alighting times are unknown).

The primary interest of this study is the variability of daily mobility patterns over the year. Daily mobility patterns are highly expected to be different for weekdays and weekends, so variability should be investigated within similar groups of days. The proposed methodology allows unsupervised identification of these groups by the introduction of a distance metric between daily mobility patterns and the application of a clustering algorithm. This approach makes the methodology more flexible, comparatively to predefined grouping criteria. The distance metric is based on the minimum-cost bipartite matching algorithm and described in Section 2; for clustering the *k*-means and hierarchical agglomerative algorithms were arbitrary chosen (further results relate to the *k*-means clustering). Number of clusters were identified on the of the gap statistics and the average silhouette width (Figure 7).

The two-cluster solution is fairly trivial – daily patterns for weekends and public holidays are recognised and grouped into one of the clusters. The three-cluster solution is more informative, so further only this solution is discussed.

Table 4 includes distribution of weekdays over the three-cluster solution.

Cluster 1 contains only weekends and public holidays, including local events (Riga City festival and Folklore festival). Clusters 2 and 3 consist of weekdays of two patterns. For discovering background reasons for splitting weekdays into two clusters, the distribution of cluster components over the seasons is considered (Table 5).

The first pattern is more typical for the summer season (43 of 52 summer weekdays belong to this cluster), while for other seasons cluster volumes are similar. One possible explanation of this fact is an effect of longer daylight time and good weather. During summer, this is more typical to visit natural attractors like sea and river beaches, forests, and smallholdings after or before regular working hours. Deeper analysis of differences between Cluster 2 and 3 days is required for better understanding of mobility patterns and fitting the PT supply.

Finally, the moving average technique is applied for analysis of mobility trends within-clusters (Figure 8).

Overall within-cluster distance has the lowest values for Cluster 1 (weekends/public holidays) and the highest values for Cluster 2 (typical for non-summer weekdays). In addition, higher variability within all clusters is observed during summer season. The higher variability that observed for all clusters at summer months are fairly explainable by tourists flows and vacations. Frequently,

Table 4. Distribution of weekdays between the clusters

Cluster number	Cluster given name	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1	Weekends and public holidays	6	2	1	0	2	43	42
2	Weekdays, pattern 1	12	19	23	18	16	0	0
3	Weekdays, pattern 2	27	25	22	29	27	1	0

Table 5. Distribution of weekday cluster components over the seasons

Cluster number	Cluster given name	Winter (December–February)	Spring (March–May)	Summer (June–August)	Autumn (September–November)
2	Weekdays, pattern 1	25	27	9	27
3	Weekdays, pattern 2	31	34	43	23

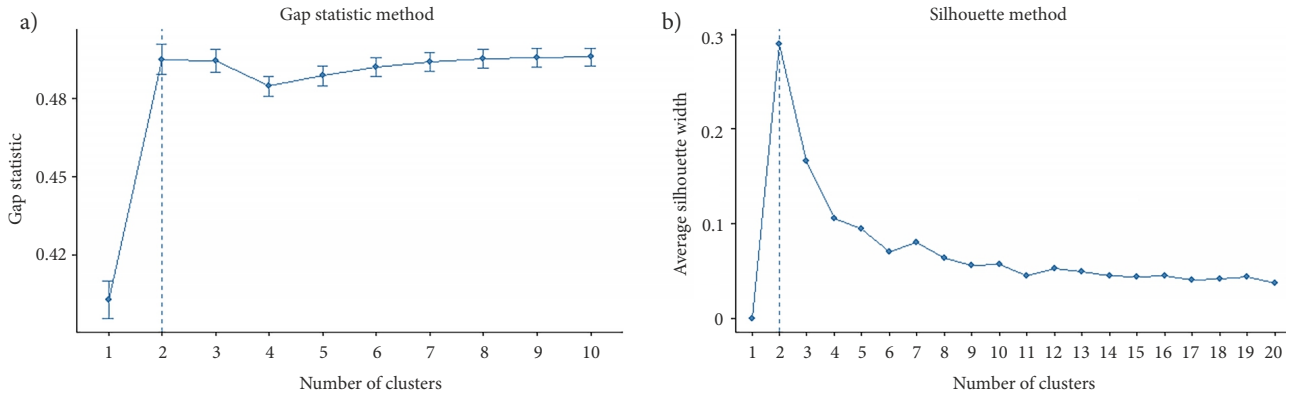

Figure 7. Gap statistics and average silhouette width charts for mobility patterns clustering

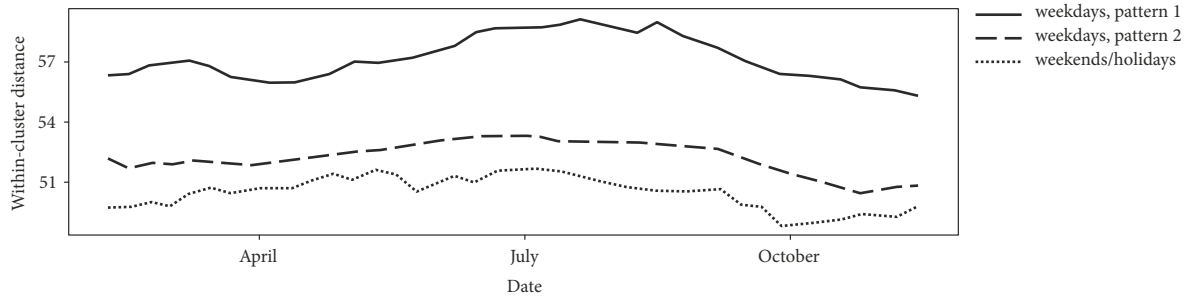

Figure 8. Variability of mobility pattern clusters (within-cluster distances, smoothed by moving average)

the most important issues of PT demand are associated with anomalies in its dynamics. For example, there is a higher variation in Cluster 2 (weekdays) during February and March, which do not have straightforward reasons. A one-year dataset is not sufficient for confident conclusions on anomalies or trend breaks, but it can be stated that ongoing monitoring of corresponding dynamics will allow decision makers to observe effects of strategic decisions like introducing of new regulations toward green transport and sustainable city development.

Reproducibility of the study. Although the case study data set is closed by the confidentiality agreement, implemented routines (*R* source codes) for all proposed algorithms and procedures were publicly provided at *https://github.com/DmitryPavlyuk/postdoc?Transport2020*. Due to a high similarity of smart card data sets for cities over the world and high availability of scheduling data in GTFS format, these routines could be helpful for reproducing the case study in other geographical settings.

4. Discussion

The added value of this study splits into practical benefits for the PT system of Riga and methodological advances of smart card data analysis. Most important practical results include:

- obtained daily mobility patterns that can be used for city-level PT OD matrix estimation that is useful in many applied areas, e.g., dynamic passenger flow assignment models;
- mobility pattern-based clustering of days allows effective comparison and flexible tuning of the PT system for different days of a week (weekdays, weekends), public holidays, extreme weather conditions, and large events;
- dynamics of mobility patterns allows estimating the effect of implementing changes (e.g., fare increase or road maintenance) and demand forecasting for user-focused development of PT system.

Summarising the case study results, its potential benefits at tactical, operational, and strategic levels should be underlined. At the tactical level, obtained mobility patterns could be used for PT service adjustment. The Riga PT operator keeps separate schedules for weekdays and weekends and adjusts schedules on a seasonal basis and for large events. Information about mobility patterns allows them better tuning of schedules to match day-specific user needs. At the operational level, the information is useful for obtaining precise performance indicators on a transit network, e.g., schedule adherence and consistency of fares. Following the proposed methodology, actual mobility patterns could be estimated at the end of a day and compared with the expected ones; significant deviation of actual and expected patters could be considered as an informative signal about potential operational problems. Long-term dynamics of mobility patterns are the most beneficial at the strategic level: better understanding of user behaviour and forecasting of its dynamics allows advanced improvements of the PT system, its extension and adaptation to passenger needs.

In addition to these benefits, the added value of this study includes the methodological component: the developed methodology allows providing the dynamic information about mobility patterns and their fluctuation over space and time almost in real-time. The methodology incorporates estimation of individual trip vectors (origin and destination spatial points), clustering of trip vectors into spatiotemporal mobility patterns, and further analysis of mobility patterns' dynamics. As smart card data are easily available, the methodology provides an easy and cheap tool for observing and managing of PT demand fluctuation on a daily basis.

The case study, presented in this paper, does not specify and utilise any clusters of passengers. The smart card database contains information about a card type (one-hour tickets, daily or monthly tickets, subsidised tickets), and the information can be even more enhanced with social and demographic data of personalised tickets. This information can be naturally used for passenger clustering and deeper analysis of their PT travel behaviour and changes of mobility patterns.

Although application of the composed methodology was successful and lead to beneficial results for our case study, several problems, which should be addressed by involved parties, were identified. Firstly, this is necessary to implement data fusion of scheduling and validation databases for obtaining spatiotemporal information. This step introduces new levels of uncertainty into the prepared data and could affect results and conclusions. At the same time, all PT vehicles in modern cities are equipped with

GPS sensors, which allow adding enhancing every smart card validation or, at least, every executed PT stop with geographical information. Such integration of on-board PT equipment is technically simple and not costly, so smart card operators are recommended to introduce this. This integration will be beneficial not only for analysis of mobility patterns, but also will help with implementation of other modern PT services like online information about loading of PT vehicles and a provided level of service. Secondly, the composed methodology includes several stages: from identification of boarding location to construction of mobility patterns. Many existing studies are focused on method development for a selected stage (e.g., identification of boarding stop locations), without its links to other, more general problems. Although development of such algorithms is beneficial, we recommend to consider their efficiency not within a scope of one stage, but as a stage of a specific practical problem solving. For example, an algorithm that provides better identification of an alighting stop in average (have a higher accuracy value) will be less efficient as a stage of a practical problem solving due to a bias in incorrectly identified locations or its strict requirement for a precise boarding location. Thirdly, it was demonstrated that the methodology of PT demand estimation could be flexible and data-driven, and does not necessarily require definition of TAZ as an input. This user-centred approach to zoning, based on real PT passenger flows and travel demand, seems to be more dynamic and informative for stakeholders and does not require a costly manual definition and updating process.

Conclusion and future work

The paper is devoted to a case study of spatiotemporal aspects of PT demand. The composed methodology of the study includes several stages, namely: fusion of information from the scheduling and smart card databases; identification of boarding locations; estimation of mobility vectors, based on the trip chain approach; construction of mobility patterns by clustering mobility vectors; clustering of dates on the base of a custom mobility pattern similarity definition; analysis of spatiotemporal mobility pattern dynamics within similar date clusters over the research period. The composed methodology was implemented as a publicly available software and allowed estimation and analysis of daily mobility patterns, which is beneficial at tactical, operational, and strategic levels.

There are a lot of opportunities for further methodological enhancements. The share of trip vectors, successfully recognised from smart card validation data, is fairly low (42…43%) and could be improved by integrating smart card and scheduling databases with PT vehicles' location information; estimating of destinations for one-time smart cards; and extracting regular routes of individual passengers. The proposed methodology highly requires an additional attention to validation of obtaining practical results using other sources of PT demand information like passenger surveys and observational studies. In addition, discovered mobility patterns and their dynamics require a deeper analysis for providing recommendations to PT operators and authorities. All mentioned issues are considered as limitations of the current study and as future steps to its practical application in Riga. Finally, taking into account the high similarity of smart card data sets for cities over the world, the proposed methodology and its software implementation can be adopted for analysis of spatiotemporal PT demand in other cities.

Acknowledgements

The authors are grateful to the Riga (Latvia) smart card operator *Rīgas Karte Ltd.* (*https://rigaskarte.lv*), for sharing the data.

Funding

Dmitry Pavlyuk and Nadežda Spiridovska were financially supported by the specific support objective activity 1.1.1.2. "Post-Doctoral Research Aid" (Project No 1.1.1.2/16/I/001) of the Republic of Latvia, funded by the European Regional Development Fund.

Dmitry Pavlyuk's research project No 1.1.1.2/VIAA/1/16/112 "Spatiotemporal urban traffic modelling using big data".

Nadežda Spiridovska's research project No 1.1.1.2/VIAA/1/16/075 "Non-traditional regression models in transport modelling".

References

Alsger, A. A.; Tavassoli, A.; Hickman, M.; Mesbah, M. 2018. Variation of transit demand based on smart card data, in *Transportation Research Board 97th Annual Meeting*, 7–11 January 2018, Washington, DC, US, 1–24.

Barry, J. J.; Freimer, R.; Slavin, H. 2009. Use of entry-only automatic fare collection data to estimate linked transit trips in New York City, *Transportation Research Record: Journal of the Transportation Research Board* 2112: 53–61. https://doi.org/10.3141/2112-07

Briand, A.-S.; Côme, E.; Trépanier, M.; Oukhellou, L. 2017. Analyzing year-to-year changes in public transport passenger behaviour using smart card data, *Transportation Research Part C: Emerging Technologies* 79: 274–289. https://doi.org/10.1016/j.trc.2017.03.021

Cats, O.; Wang, Q.; Zhao, Y. 2015. Identification and classification of public transport activity centres in Stockholm using passenger flows data, *Journal of Transport Geography* 48: 10–22. https://doi.org/10.1016/j.jtrangeo.2015.08.005

Chen, X.; Wang, Y.; Tang, J.; Dai, Z.; Ma, X. 2020. Examining regional mobility patterns of public transit and automobile users based on the smart card and mobile Internet data: a case study of Chengdu, China, *IET Intelligent Transport Systems* 14(1): 45–55. https://doi.org/10.1049/iet-its.2019.0333

Chen, Z.; Fan, W. 2018. Extracting bus transit boarding stop information using smart card transaction data, *Journal of Modern Transportation* 26(3): 209–219. https://doi.org/10.1007/s40534-018-0165-y

Deschaintres, E.; Morency, C.; Trépanier, M. 2019. Analyzing transit user behavior with 51 weeks of smart card data, *Transportation Research Record: Journal of the Transportation Research Board* 2673(6): 33–45. https://doi.org/10.1177/0361198119834917

EC. 2013. *Annex 1: a Concept for Sustainable Urban Mobility Plans to the Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions*. COM(2013) 913 Final. European Commission (EC). 5 p. Available from Internet: https://eur-lex.europa.eu/resource.html?uri=cellar:82155e82-67ca-11e3-a7e4-01aa75ed71a1.0011.02/DOC_4&format=PDF

Egu, O.; Bonnel, P. 2020. Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. A case study in Lyon, *Travel Behaviour and Society* 19: 112–123. https://doi.org/10.1016/j.tbs.2019.12.003

El Mahrsi, M. K.; Côme, E.; Oukhellou, L.; Verleysen, M. 2017. Clustering smart card data for urban mobility analysis, *IEEE Transactions on Intelligent Transportation Systems* 18(3): 712–728. https://doi.org/10.1109/TITS.2016.2600515

Faroqi, H.; Mesbah, M.; Kim, J. 2018. Applications of transit smart cards beyond a fare collection tool: a literature review, *Advances in Transportation Studies* 45: 107–122.

Gentile, G.; Noekel, K. 2016. *Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems: COST Action TU1004* (*TransITS*). Springer. 641 p. https://doi.org/10.1007/978-3-319-25082-3

Goulet-Langlois, G.; Koutsopoulos, H. N.; Zhao, J. 2016. Inferring patterns in the multi-week activity sequences of public transport users, *Transportation Research Part C: Emerging Technologies* 64: 1–16. https://doi.org/10.1016/j.trc.2015.12.012

Goulet-Langlois, G.; Koutsopoulos, H. N.; Zhao, Z.; Zhao, J. 2018. Measuring regularity of individual travel patterns, *IEEE Transactions on Intelligent Transportation Systems* 19(5): 1583–1592. https://doi.org/10.1109/TITS.2017.2728704

Han, G.; Sohn, K. 2016. Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model, *Transportation Research Part B: Methodological* 83: 121–135. https://doi.org/10.1016/j.trb.2015.11.015

He, L.; Agard, B.; Trépanier, M. 2020. A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method, *Transportmetrica A: Transport Science* 16(1): 56–75. https://doi.org/10.1080/23249935.2018.1479722

Kieu, L. M.; Bhaskar, A.; Chung, E. 2015. Passenger segmentation using smart card data, *IEEE Transactions on Intelligent Transportation Systems* 16(3): 1537–1548. https://doi.org/10.1109/TITS.2014.2368998

Kumar, P.; Khani, A.; He, Q. 2018. A robust method for estimating transit passenger trajectories using automated data, *Transportation Research Part C: Emerging Technologies* 95: 731–747. https://doi.org/10.1016/j.trc.2018.08.006

Long, Y.; Thill, J.-C. 2015. Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing, *Computers, Environment and Urban Systems* 53: 19–35. https://doi.org/10.1016/j.compenvurbsys.2015.02.005

Ma, X.; Wu, Y.-J.; Wang, Y.; Chen, F.; Liu, J. 2013. Mining smart card data for transit riders' travel patterns, *Transportation Research Part C: Emerging Technologies* 36: 1–12. https://doi.org/10.1016/j.trc.2013.07.010

Manley, E.; Zhong, C.; Batty, M. 2018. Spatiotemporal variation in travel regularity through transit user profiling, *Transportation* 45(3): 703–732. https://doi.org/10.1007/s11116-016-9747-x

Morency, C.; Trépanier, M.; Agard, B. 2007. Measuring transit use variability with smart-card data, *Transport Policy* 14(3): 193–203. https://doi.org/10.1016/j.tranpol.2007.01.001

Munizaga, M. A.; Palma, C. 2012. Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from Santiago, Chile, *Transportation Research Part C: Emerging Technologies* 24: 9–18. https://doi.org/10.1016/j.trc.2012.01.007

OMD. 2019. *Open Mobility Data*. Public Database. Available from Internet: https://openmobilitydata.org

Pelletier, M.-P.; Trépanier, M.; Morency, C. 2011. Smart card data use in public transit: a literature review, *Transportation Research Part C: Emerging Technologies* 19(4): 557–568. https://doi.org/10.1016/j.trc.2010.12.003

Qi, G.; Huang, A.; Guan, W.; Fan, L. 2019. Analysis and prediction of regional mobility patterns of bus travellers using smart card data and points of interest data, *IEEE Transactions on Intelligent Transportation Systems* 20(4): 1197–1214. https://doi.org/10.1109/TITS.2018.2840122

Rupprecht, S.; Brand, L.; Böhler-Baedeker, S.; Brunner, L. M. 2019. *Guidelines for Developing and Implementing a Sustainable Urban Mobility Plan*. 2nd Edition. European Platform on Sustainable Urban Mobility Plans 166 p. Available from Internet: https://www.eltis.org/sites/default/files/sump-guidelines-2019_mediumres.pdf

Tao, S.; Rohde, D.; Corcoran, J. 2014. Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap, *Journal of Transport Geography* 41: 21–36. https://doi.org/10.1016/j.jtrangeo.2014.08.006

Trépanier, M.; Tranchant, N.; Chapleau, R. 2007. Individual trip destination estimation in a transit smart card automated fare collection system, *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations* 11(1): 1–14. https://doi.org/10.1080/15472450601122256

Wang, W.; Attanucci, J.; Wilson, N. H. M. 2011. Bus passenger origin-destination estimation and related analyses using automated data collection systems, *Journal of Public Transportation* 14(4): 131–150. https://doi.org/10.5038/2375-0901.14.4.7

Wei, S.; Yuan, J.; Qiu, Y.; Luan, X.; Han, S.; Zhou, W.; Xu, C. 2017. Exploring the potential of open big data from ticketing websites to characterize travel patterns within the Chinese high-speed rail system, *PLoS ONE* 12(6): e0178023. https://doi.org/10.1371/journal.pone.0178023

Welch, T. F.; Widita, A. 2019. Big data in public transportation: a review of sources and methods, *Transport Reviews* 39(6): 795–818. https://doi.org/10.1080/01441647.2019.1616849

Xu, Z.; Cui, G.; Zhong, M.; Wang, X. 2019. Anomalous urban mobility pattern detection based on GPS trajectories and POI data, *ISPRS International Journal of Geo-Information* 8(7): 308. https://doi.org/10.3390/ijgi8070308

Zhong, C.; Manley, E.; Arisona, S. M.; Batty, M.; Schmitt, G. 2015. Measuring variability of mobility patterns from multiday smart-card data, *Journal of Computational Science* 9: 125–130. https://doi.org/10.1016/j.jocs.2015.04.021