Data Science for Transportation (2023) 5:20
https://doi.org/10.1007/s42421-023-00082-x

RESEARCH

Extracting Spatiotemporal Bus Passenger Trip Typologies from Noisy Mobile Ticketing Boarding Data

Mohammed Abdalazeem[1] · **Jimi Oke**[1]

Received: 28 November 2022 / Revised: 28 June 2023 / Accepted: 14 August 2023 / Published online: 14 September 2023

Abstract

We present a framework for extracting spatiotemporal trip typologies using noisy mobile ticketing boarding data sampled from passengers in a bus network. Our case study was the Pioneer Valley Transit Authority in Massachusetts. We first used a greedy approach to infer bus boarding stops. Next, we calculated the multi-dimensional dissimilarity of passenger activation time series using the AWarp alignment algorithm for sparse time series. We then employed hierarchical clustering to discover the spatiotemporal patterns, resulting in four distinct trip pattern typologies. We analyzed the typologies, based on trip length and duration, seasonality and other temporal distributions, spatial distributions, and faretype. Three typologies were linked to regular commuters, distinguished by boarding time or transfer tendency. The fourth typology was primarily associated with leisure or other activities. Our typology method provides valuable passenger behavioral insights and can facilitate demand estimation by planners. Further, we demonstrate a potential for decision-making support for other regional transit authorities with limited passenger data availability.

Keywords Spatiotemporal patterns · Clustering · Mobile ticketing · Dynamic time warping · Public transportation · Typologies

Introduction

Public transportation ridership in the United States has been in decline over the past decade. Between 2014 and 2018, ridership decreased by 7.5% (O'Toole 2018). The recent COVID-19 pandemic further highlighted the vulnerability of public transportation to disruptive events, as usage was severely curtailed across the US, and even worldwide (Liu et al. 2020). In the US, ridership dropped to 20% of pre-pandemic levels in April 2020 but has since rebounded to more than 70% of pre-pandemic levels (Kahana and Dickens 2023). With the continued rise of new modes and mobility services, as well as the increasingly imminent threats of disasters or extreme events like the COVID-19 pandemic, there is an imperative to understand passenger behavior and travel patterns. This understanding is crucial to guiding planning and response efforts, consistently meeting ridership demand, and promoting public transportation usage.

Passengers in public transportation networks have unique travel patterns and characteristics that transit agencies have typically captured using surveys. However, survey data often suffer from bias due to small sample sizes and can quickly become outdated. Conducting surveys can also be resource-intensive (Ben-Akiva and Morikawa 1989; Cui 2006). Automated fare collection systems (AFC) offer a more efficient alternative, collecting extensive data including boarding location, time, faretype, and demographics (Sun and Xu 2012). Stored daily and anonymized, these data allow for large-scale tracking of passenger activity. However, transit agencies often lack the tools to collect and effectively analyze this wealth of information, hindering their ability to adapt to changing demands and constraints (Zhang et al. 2018; Chen et al. 2020; Zhang 2022).

Notably, smaller transit agencies often lack access to the extensive data sources available to larger ones. Thus, we present an analytical framework that leverages noisy boarding-only mobile ticketing data for travel pattern analysis, providing valuable insights into passenger behavior and network

✉ Mohammed Abdalazeem
mamohammed@umass.edu

Jimi Oke
jboke@umass.edu

[1] Department of Civil and Environmental Engineering, University of Massachusetts Amherst, 130 Natural Resources Rd, Amherst, MA 01003, USA

usage at a reduced cost. Specifically, we utilize the dynamic time warping algorithm along with hierarchical clustering to infer and analyze the spatiotemporal trip pattern typologies of passengers in the Pioneer Valley Transit Authority (PVTA) bus system. This framework enables planners to better comprehend passenger behavior, make informed decisions, and optimize resource allocation. Furthermore, we demonstrate how the inclusion of demographic and faretype data can enhance equitable planning.

Related Work

In the 1970 s and 1980 s, many surveys highlighted the shortcomings of transit models focusing on aggregate demand at the zonal level rather than diverse person-level day-to-day demand patterns (Manley et al. 2018; Jones and Clarke 1988; Pas and Koppelman 1986; Pas 1987). Additionally, several studies have indicated the richness, complexity, and importance of capturing person-level patterns (Manley et al. 2018; Hanson and Huff 1986, 1988; Zhong et al. 2015; Primerano et al. 2008), which can be obtained by analyzing user activities spatiotemporally, i.e., through time and space (Manley et al. 2018).

Over the past few decades, automated fare collection (AFC) systems, particularly those based on smart cards and mobile devices, have revolutionized the analysis of travel patterns in public transportation by collecting detailed high-resolution spatiotemporal data (Manley et al. 2018). AFC systems have been increasingly used for inferring certain aspects of passenger behavior, such as station demand, transfer locations, and desired destinations (Mohammed and Oke 2023). In particular, those patterns have been inferred in the form of origin–destination (OD) matrices based on a variety of data sources, including surveys, mobility traces, and observed counts, using various methods, such as maximum entropy (Ait-Ali and Eliasson 2019; Ge and Fukuda 2016), Bayesian inference (Hazelton 2010), maximum likelihood (Navick and Furth 1994), and optimization models that identify transfers and approximate network-level OD flows (Liu et al. 2021). Trip chaining is the most commonly used method for inferring passenger behavior through OD matrices. It primarily utilizes AFC data to support planning and operations by leveraging economically collected, high-resolution data. This method helps in estimating the number of passengers boarding and alighting at each bus stop, either at the route level (Alsger et al. 2015; Wang 2010), or across the entire network (Cui 2006; Gordon 2012).

Other aspects of user behavior, such as passenger spatiotemporal patterns, are not collected by AFC systems and cannot typically be derived from OD matrices. To address this gap, several sophisticated techniques have been developed over the years for analyzing and extracting meaningful insights from passenger spatiotemporal patterns. These techniques encompass a range of methodologies, including pattern prediction, change detection, and clustering, and often employ advanced machine learning algorithms to infer the underlying trends and characteristics in the data (Shi and Pun-Cheng 2019). Examples of these spatiotemporal pattern detection methods in the literature include: dynamic aggregation of movement data (Rinzivillo et al. 2008), deep embedded clustering (Asadi and Regan 2019), Moran's I method of spatial autocorrelation with kernel density functions (Prasannakumar et al. 2011), support vector regression (Asif et al. 2014), negative binomial regression (Hochmair 2016), temporal Fuzzy C-Means (FCM) clustering (Sun et al. 2018; Zhong and Sun 2022), and K-means clustering (Song et al. 2019; Sanaullah et al. 2021; Chen et al. 2019). Generally, clustering methods are some of the most widely used (Shi and Pun-Cheng 2019).

Spatiotemporal clustering involves the grouping of trajectories based on similarities in their features. Clustering methods include distance-based, density-based, visual-aided, and micro clustering methods (Kisilevich et al. 2010; Manley et al. 2018; Giannotti et al. 2007; Shi and Pun-Cheng 2019). Similarities are computed using distance metrics based on different trajectory properties, such as origins, destinations, start and end times, or even routes (Rinzivillo et al. 2008). Spatiotemporal distance algorithms range from the very simple ones that compute the average Euclidean distances between coordinates and the absolute time between each trajectory's start and end (such as common source and destination), to the highly complex methods that incorporate a comparison of routes taken as well. However, the high dimensionality of the data and the sparsity of the time series can lead to challenges, such as in computational performance and accuracy. In spite of these issues, clustering still proves popular due to its simplicity and its ability to yield highly interpretable results.

In public transportation applications, specifically, some efforts to analyze travel patterns simply compute statistics on existing data using predefined passenger groups (Shi et al. 2020; Nishiuchi et al. 2013; Shao et al. 2019). Typically, demographics are used. These approaches are less time-consuming and can still yield important behavioral insights, especially when relevant variables are available in the data. Other implementations have used automated fare collection (AFC) data along with clustering methods to extract spatial and temporal features of passengers for pattern analysis. For instance, some studies clustered spatial and temporal patterns using Poisson and Gaussian mixture models (El Mahrsi et al. 2017; Briand et al. 2017), while others have employed the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to detect historical travel patterns with rough-set theory (Ma et al. 2013) and to discern differences between regular and irregular travel (Manley

et al. 2018), and others have used unsupervised learning algorithms to study commuting trips based on GPS data (Costa et al. 2023). Additionally, K-means clustering, which necessitates specifying the number of clusters, has been employed by some researchers (Zhao et al. 2014; Inmook 2019; Zhao et al. 2017; Ma et al. 2013; Agard et al. 2013). Recent developments include the application of hierarchical clustering with cross-correlation distance to time-series data for classifying daily transactions of public transit smart card users (He et al. 2020), the adoption of clustering techniques to analyze smart card data focusing on station-oriented commute space patterns (Yong et al. 2021), and the introduction of a novel modulated spatiotemporal clustering tool that analyzes smart card data by controlling the relative influence of space and time (Decouvelaere et al. 2022).

The aforementioned methods typically require extensive data collection and significant computational resources, which can be expensive and time-consuming. Furthermore, some of these methods may not be suitable for regional bus networks, where data availability and the network scale can pose unique challenges. Thus, there is a need for methods that are cost-effective and efficient in performing travel pattern analyses in such networks.

We address this gap by developing a framework for analyzing travel patterns in a regional bus network based on clustering anonymized passenger boarding activations on a mobile platform. Our contribution is significant in two ways. First, we have a minimal number of variables compared to the more detailed information available from smart card datasets used in larger bus systems. Second, we applied a recent innovation in computing dissimilarity between sparse time series observations—the AWarp algorithm (Mueen et al. 2018). Our approach demonstrates the viability of clustering techniques for transit agencies that have noisy mobile boarding ticketing data such as the one used here.

Data and Methods

In this section, we present the case study network and the data used for our analyses. We then provide the approaches used to infer the origins of trips, including the greedy matching approach and the DBSCAN spatial clustering approach, and compare them to assess their effectiveness. After identifying boarding stops, we construct passenger activation time series, encompassing the chronological sequence of passenger activations, including boarding locations and timestamps. These data are then subjected to dynamic time warping and hierarchical clustering for the identification of distinct passenger typologies. The entire framework, from data input and processing to typology identification, is shown in Fig. 1.

Case Study Area

The Pioneer Valley Transit Authority (PVTA) bus network in the western Pioneer Valley region of the US state of Massachusetts served as our case study. Founded in 1974, the PVTA is the state's largest regional transit authority, serving 24 communities in Hampden and Hampshire counties. Its service area spans over 600 square miles and covers a population of about 600000. The network includes 186 buses operating on 36 routes with 1811 bus stops (PVTA 2023).

The PVTA employs an entry-only automated fare collection (AFC) system with a flat fare structure, facilitated by a mobile ticketing application introduced in July 2020. At the time of the study, only a small percentage of PVTA passengers (10–12%) were using this system. The system requires passengers to activate their tickets around the time of boarding, after which the driver visually verifies them. Passengers can also make unlimited transfers within the ticket's validity period using the application. Notably, passengers may activate their tickets more than once before boarding, leading to multiple records in the dataset for a single boarding. These activations may occur at different locations near the bus stop, and as such, the recorded activation locations may not always align with the actual bus stops. This imprecision, coupled with multiple activations, complicates the task of accurately pinpointing boarding locations. The dataset used in this study is anonymized and comprises 3201 unique passengers with a total of 257000 passenger activations from July 20, 2020, to March 28, 2021. It includes data fields such as user ID, timestamp, coordinates, faretype, and an indicator for elderly or disabled passengers (E&D).

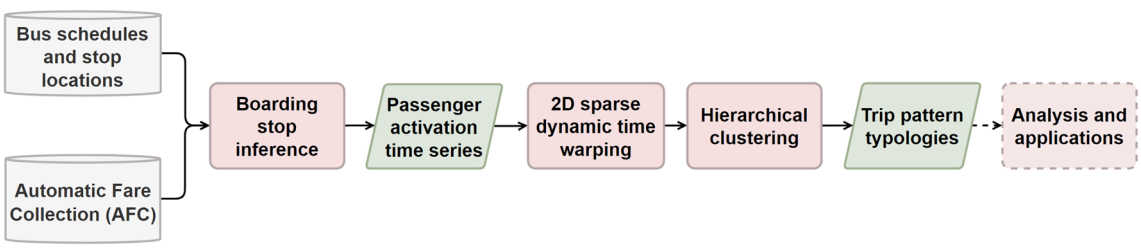

Fig. 1 Methodological framework for the spatiotemporal analysis of bus ridership

Figure 2 illustrates the distribution of activation locations in juxtaposition to bus stops. Notably, the activations are not just dispersed over an extensive geographic area, but also appear in locations with no adjacent bus stops. Where activations are close to bus stops, they are often erratically scattered, at times closer to multiple stops or nearer to an unintended stop rather than the actual one. The resulting noise and duplication require algorithms for cleaning and accurately assigning activations with the corresponding bus stops. Additionally, the hourly fluctuations in passenger activations, which follow the anticipated morning and afternoon peaks, are depicted in Fig. 3.

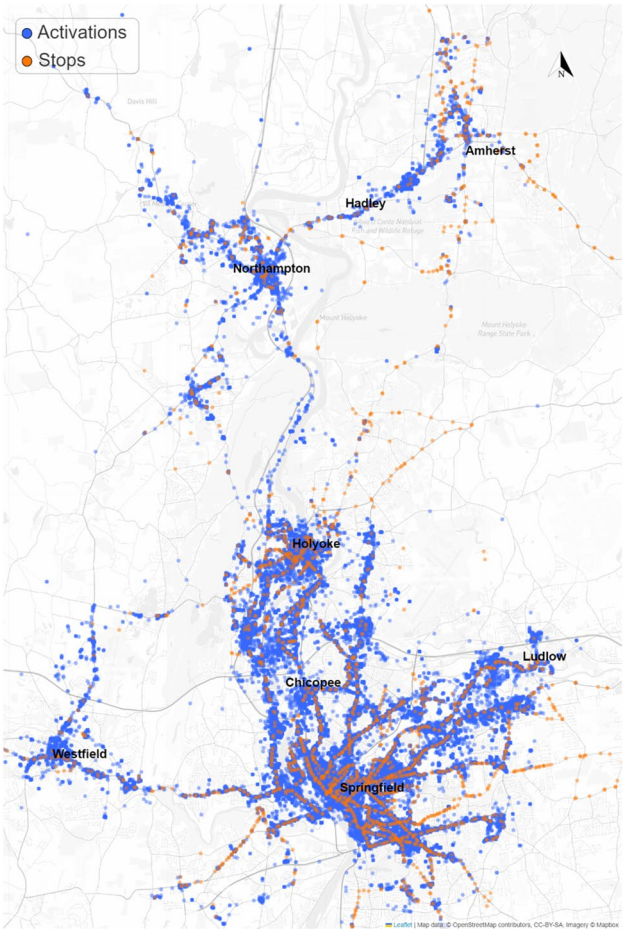

Fig. 2 Map of bus stops and user activations in the Pioneer Valley Transit Authority network from July 2020 through March 2021

Boarding Inference

We began by inferring the boarding locations of trips from the noisy mobile ticket activation data. This involved assigning individual activations to bus stops since the data did not include information on origin stops. Generally, passengers activate their tickets around the time they board the bus, marking the start of a new trip segment. However, tickets can be activated at any time and place, which may result in false activations. Our goal was to use the recorded locations of activations to determine the boarding stops, filtering out false activations.

We evaluated two methods to assign activations to stops: (a) spatial clustering-based assignment, and (b) greedy distance-based assignment. The spatial clustering method groups activations based on their density, while the greedy distance-based method assigns each activation to nearest stop, which is simpler and more computationally efficient. We used the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm (Ester et al. 1996) for spatial clustering, which identifies clusters of nearby observations. DBSCAN requires two parameters: a minimum number of points (*minPoints*) to define a region as a cluster, and a distance threshold (ϵ) beyond which observations are not considered neighbors. We conducted a grid search over a range of values, $minPoints \in \{1, 2, 3, \ldots, 10\}$ and $\epsilon \in \{5, 10, 15, 20, 25\}$ m, to find the optimal values, $minPoints^* = 2$ and $\epsilon^* = 15$ m, and ran the clustering algorithm on the activations. After excluding noise points (observations not in any cluster), we matched each cluster to the nearest stop within a 100 ms to avoid incorrect assignments. Each activation was assigned to the stop corresponding to its cluster. In the greedy method, we first computed the distance from each activation location to every stop, then directly assigned each activation to the nearest stop within a radius $r = 100$ m.

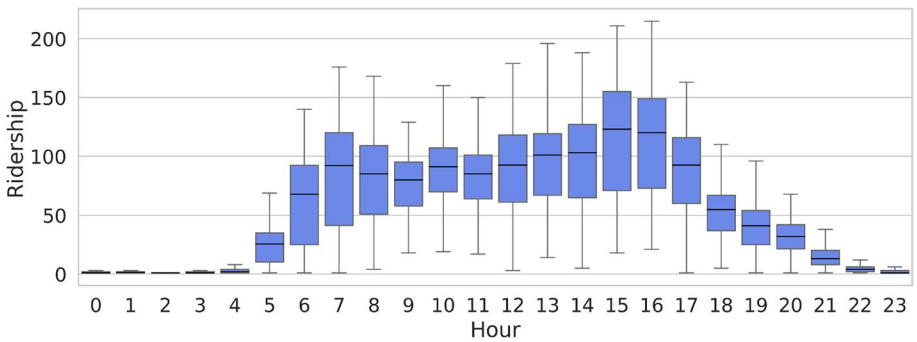

Fig. 3 Hourly distribution of passenger activations from July 2020 through March 2021

Given the absence of ground truth data for validation, we compared the methods based on computation time, the percentage of activations successfully assigned to an active stop, and the ratio of activation-matched stops to all stops in the network. These metrics were calculated for the full dataset over 8 months and were also averaged by month, week, and day. The comparison is shown in Table 1.

The results indicated that the greedy method consistently assigned a higher percentage of activations to active stops compared to DBSCAN, particularly when data were provided at smaller time scales. This is because DBSCAN requires a certain density of points to form clusters and removes points not meeting this threshold as noise. In contrast, the greedy method uses all points and assigns them to stops based on proximity. Moreover, the stop matching rate was significantly higher using the greedy method. This is because, in DBSCAN, points near multiple stops may be grouped into one cluster, leading to assignment to only one stop. We also observed that DBSCAN was faster on small time scales, but the greedy method was faster on larger time scales, making it more suitable for our dataset. Consequently, we proceeded with the greedy assignment method to infer boarding locations.

Passenger Activation Time Series Preparation

We obtained the noisy activation data from the PVTA, which contained ridership data for the 8-month period from July 2020 to March 2021, with the goal of using them to obtain groups of passengers with similar mobility patterns. To render the data suitable for spatiotemporal time series extraction, we created a longitudinal dataset containing observations of each individual's activation locations over the study time period. We aggregated the data at the 10-min level, meaning only one activation is recorded if multiple activations occur within a 10-min window. This choice of a 10-min scale was made based on a balance between granularity and computational efficiency. A finer resolution would be computationally expensive and may not provide substantial additional insights, while a coarser resolution might obscure important patterns. Thus, a 10-min scale strikes a balance between capturing the nuances in passengers' travel behaviors and being within a reasonable spatiotemporal radius for more than one activation point to be considered one, while maintaining reasonable computational efficiency. Furthermore, it is within a suitable spatiotemporal radius to consider multiple activation points as part of the same trip.

The result was a dataset that consists of individual separate passenger records that can be clustered. We represented this dataset as a data tensor, $\mathcal{Y} \in \mathbb{R}^{N \times D \times T}$, where $N = 3201$ is the number of passengers, $D = 2$ are the dimensions for latitude and longitude, and $T = 36684$ is the number of 10-min time intervals in the study period (from July 20, 2020, to March 28, 2021). The dataset starts with the first date a passenger activated a ticket. Each subsequent cell contains either the longitude or latitude of the first activation within a 10-minute period, or a zero if no ticket was activated. A snapshot of the time series table is provided in Table 2.

Dynamic Time Warping

Clustering requires pairwise dissimilarities of the observations, which in this case are passenger time series. To compute the similarity between time series, we used the dynamic time warping (DTW) algorithm (Sakoe and Chiba 1978; Salvador and Chan 2007), which was originally developed as a method for speech recognition. The similarity is computed by finding the optimal warping curve ϕ_k that minimizes the weighted average distortion between corresponding elements of a pair of two time-series vectors $\boldsymbol{y}_1 \in \mathbb{R}^{T_1}$ (known as the "reference") and $\boldsymbol{y}_2 \in \mathbb{R}^{T_2}$ ("query") whose lengths may vary. The warping curve ϕ_k specifies the $k = 1, \ldots, M$ mappings of elements in $\boldsymbol{y}_1$ to those in $\boldsymbol{y}_2$, where $M = \max(T_1, T_2)$. The DTW distance δ between time series $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ is thus given as:

Table 1 Performance metrics comparing the spatial clustering approach to the greedy distance-based approach for assigning passenger activations to bus stop locations in the network

Method	Time scale	Computation time (seconds)	Boarding inference rate (%)	Stop matching rate (%)
Spatial clustering	8 months	1245.0	60.9	69.4
	Month	26.5	73.3	40.7
	Week	6.0	66.6	23.8
	Day	0.7	50.8	6.0
Greedy distance-based	8 months	405.0	79.0	83.7
	Month	43.8	78.0	58.4
	Week	12.0	78.6	40.4
	Day	1.7	78.7	16.0

Table 2 A subsample of the two-dimensional time series tables representing slices of the data tensor $\mathcal{Y}$ at $d = 1$ (longitude) and $d = 2$ (latitude). Zero values indicate no activations

	Longitude ($d = 1$)				Latitude ($d = 2$)			
Time (t)	1	2	3	...	1	2	3	...
User (n)								
1	0	0	0	...	0	0	0	...
2	0	0	0	...	0	0	0	...
3	72.5600	0	0	...	42.1067	0	0	...
4	0	0	72.5252	...	0	0	42.8995	...
5	0	0	0	...	0	0	0	...
6	0	72.0009	0	...	0	42.8495	0	...
7	0	0	72.5432	...	0	0	42.1123	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3201	0	0	72.4300	...	0	0	42.5020	...

$$\delta(\boldsymbol{y}_1, \boldsymbol{y}_2) = \frac{1}{C} \sum_{k=1}^{M} c_k d(\phi_k(\boldsymbol{y}_1, \boldsymbol{y}_2)) \tag{1}$$

where C is a normalization constant, c_k is the weighting coefficient in each step and $d(\cdot)$ is a metric, such as the Euclidean distance function.

Historically used on one-dimensional data, DTW was later expanded for multiple dimensions (Shen and Chi 2021; Shokoohi-Yekta et al. 2017). A few methods exist, but we used the independent approach for multi-dimensional DTW in which pairwise distances between observations, δ_I, are computed separately in each dimension and then summed, as follows:

$$\delta_I(\boldsymbol{Y}_n, \boldsymbol{Y}_{n'}) = \sum_{d=1}^{D} \delta(\boldsymbol{y}_{n,d}, \boldsymbol{y}_{n',d}) \tag{2}$$

where $\boldsymbol{Y}_n, \boldsymbol{Y}_{n'} \in \mathbb{R}^{D \times T}$ are multivariate time series matrices representing activations of a pair of passengers n and n'.

However, DTW can be time-consuming and extremely inefficient, especially when applied to long and sparse time series data that contain mostly zeros (Mueen et al. 2018; Hwang and Gelfand 2018). Thus, we used the AWarp algorithm (implemented in MATLAB), which takes advantage of sparsity in employing a run-length encoding compression method that has been shown to drastically outperform traditional DTW in terms of computation time (Mueen et al. 2018). Additionally, the AWarp algorithm adeptly handles zeros (missing values) automatically by computing the best alignment of the time series data while accounting for temporal gaps and adjusting the similarity measure accordingly. This makes the AWarp algorithm particularly effective in processing and comparing passenger profiles without requiring explicit handling of the zeros. Ultimately, we obtained a dissimilarity matrix $\boldsymbol{\Delta} \in \mathbb{R}^{N \times N}$ of pairwise distances between the activation time series matrices $\boldsymbol{Y}_n$ and $\boldsymbol{Y}_{n'}$ of the passengers.

Hierarchical Clustering

We clustered the passenger activation time series to find the prevailing spatiotemporal patterns in the data using hierarchical agglomerative clustering (HAC) approaches (implemented with Python Scikit-learn (Cournapeau 2007)). HAC allows for an examination of the entire grouping structure. Thus, it provides a means of choosing the optimal/best number of clusters without specifying it beforehand. For our dataset, we found that Ward's linkage algorithm (also known as Ward's minimum variance method) (Ward 1963) gave the best interpretation. Ward's algorithm iteratively joins two clusters $\mathcal{C}_k$ and $\mathcal{C}'_k$ by calculating the distance between each cluster centroid and its members and minimizing the increase of the sum of squared errors (SSE). The goal is to minimize the difference between the combined SSE and the sum of the individual clusters' SSE, resulting in the optimal pair of clusters to join, $\mathcal{C}^*_{(k,k')}$. The optimization problem (Strauss and von Maltitz 2017) is specified as:

$$\mathcal{C}^*_{(k,k')} = \underset{\mathcal{C}_k, \mathcal{C}_{k'}}{\operatorname{argmin}} \frac{|\mathcal{C}_k||\mathcal{C}_{k'}|}{|\mathcal{C}_k| + |\mathcal{C}_{k'}|} \|\bar{\boldsymbol{y}}_{\mathcal{C}_k} - \bar{\boldsymbol{y}}_{\mathcal{C}_{k'}}\|^2 \tag{3}$$

where $|\mathcal{C}_k|$ and $|\mathcal{C}_{k'}|$ are the respective cluster sizes, $\|.\|^2$ denotes the squared Euclidean distance, and $\bar{\boldsymbol{y}}_{\mathcal{C}_k}$ and $\bar{\boldsymbol{y}}_{\mathcal{C}_{k'}}$ are the respective cluster centroids.

We examined two cluster quality metrics, namely: the silhouette coefficient and the Calinski–Harabasz Index, to determine the optimal number of clusters to extract from the hierarchical clustering analysis. The mean silhouette coefficient S measures how tightly each observation fits in its assigned cluster (Rousseeuw 1987). It is defined as:

$$S = \frac{1}{N} \sum_{n=1}^{N} \frac{\frac{1}{|\mathcal{C}_{k'}|} \sum_{n' \in \mathcal{C}_{k'}} \|\boldsymbol{y}_n - \boldsymbol{y}_{n'}\|^2 - \frac{1}{|\mathcal{C}_k|} \sum_{n' \in \mathcal{C}_k} \|\boldsymbol{y}_n - \boldsymbol{y}_{n'}\|^2}{\max\left(\frac{1}{|\mathcal{C}_{k'}|} \sum_{n' \in \mathcal{C}_{k'}} \|\boldsymbol{y}_n - \boldsymbol{y}_{n'}\|^2, \frac{1}{|\mathcal{C}_k|} \sum_{n' \in \mathcal{C}_k} \|\boldsymbol{y}_n - \boldsymbol{y}_{n'}\|^2\right)} \tag{4}$$

where the first term in the numerator represents the average distance of the n-th observation to all observations in the nearest neighboring cluster $\mathcal{C}_{k'}$ and the second term is the average distance of the n-th observation to all other observations within the same cluster $\mathcal{C}_k$. The values range between −1 and +1, with +1 indicating the best cluster separation. The Calinski–Harabasz Index (CH) is the ratio of the between-cluster and within-cluster variances (Caliński and Harabasz 1974). It is given by:

$$CH = \frac{(N-K)\sum_{k=1}^{K} |\mathcal{C}_k| \|\bar{\mathbf{y}}_{\mathcal{C}_k} - \bar{\mathbf{y}}_{\mathbf{\Delta}}\|^2}{(K-1)\sum_{k=1}^{K} \sum_{n \in \mathcal{C}_k} \|\mathbf{y}_n - \bar{\mathbf{y}}_{\mathcal{C}_k}\|^2} \tag{5}$$

where N is the total number of observations in the dataset, K is the number of clusters, $\mathbf{y}_n$ is observation n in cluster $\mathcal{C}_k$, and $\bar{\mathbf{y}}_{\mathbf{\Delta}}$ is the centroid of all data points considered as a single cluster.

Results and Discussion

Spatiotemporal Trip Pattern Typologies

After applying Ward's method, we obtained the dendrogram shown in Fig. 4.

We inspected the dendrogram to determine the optimal number of clusters. Three or four clusters appeared to provide the most distinct yet interpretable groups. The values for the silhouette metric and the Calinski–Harabasz (CH) Index over different cluster numbers are shown in Fig. 5.

The silhouette score is largely stable for various numbers of clusters. For the CH index, however, as the number of clusters increases, the score decreases. A significant drop in the index was observed for five clusters when compared to four clusters. Due to the need to have fewer clusters for the sake of interpretability, we found that four was the optimal number of clusters to extract. We refer to these clusters as spatiotemporal trip pattern typologies, namely T1, T2, T3,

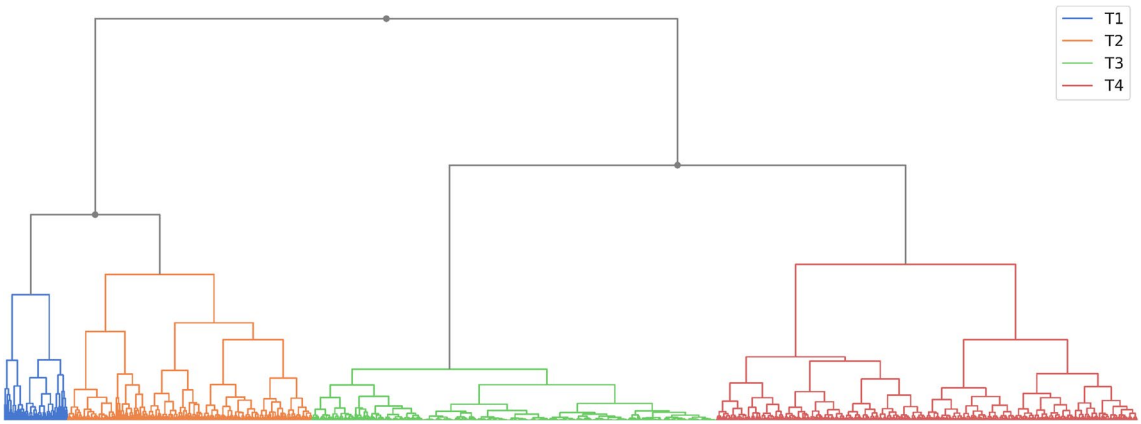

Fig. 4 Hierarchical clustering dendrogram of passenger activations trajectory similarities from July 2020 through March 2021 (using Ward's method) showing the four clusters (typologies) extracted

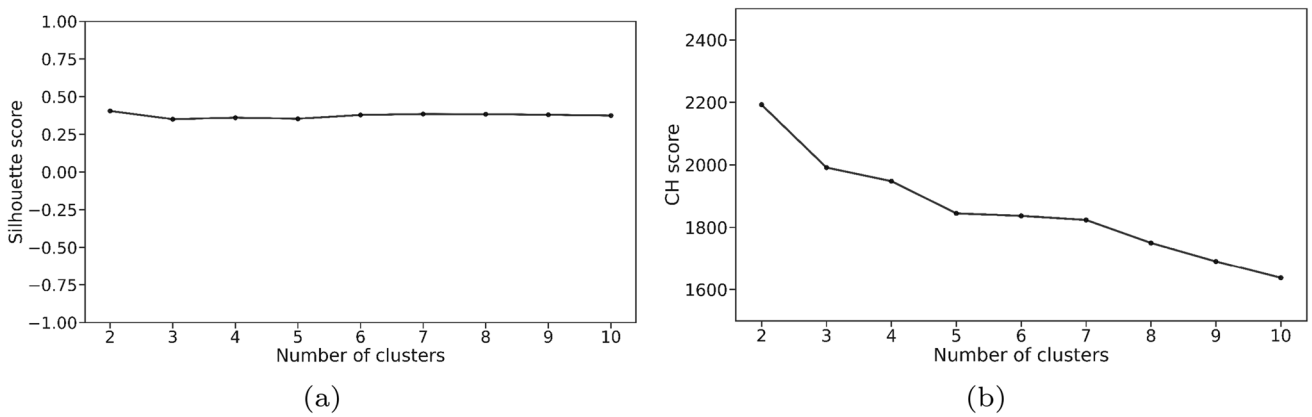

Fig. 5 Values of the (**a**) silhouette metric and (**b**) Calinski–Harabasz (CH) Index for various cluster numbers

and T4. We provide a summary of all the key characteristics of the typologies in Table 3.

We investigated the characteristics of these typologies based on the following outcomes: hourly ridership patterns, average user time intervals between activations, as well as weekday and weekend patterns. We also analyzed faretype usage and investigated trip and faretype patterns across the typologies using network maps.

Temporal Patterns

User Time Interval Between Activations

The time between each of the passenger's activations can reveal the daily average number of trips made by individual passengers. A shorter average time interval between activations for a given passenger implies that they have more bus boardings during a day, and consequently trips, compared to a passenger with a relatively longer average time interval. Furthermore, shorter time intervals may also indicate a high frequency of transfers. We show the distribution of the user average time between activations for each typology in Fig. 6. Summary statistics are given in Table 4.

In Fig. 6, we observe that T3 has a solitary peak at the 4-hour gap between activations, as demonstrated by the modal interval (Table 4). Consequently, passengers belonging to these typologies tend to re-board the bus more rapidly in comparison to passengers from other typologies. This suggests a greater transfer rate for passengers within this typology.

Observations in T1 are concentrated between 9 and 10 h, which is consistent with normal work hours. This pattern indicates that passengers from this typology mostly have no need to transfer and only use one bus to reach their destinations. It also supports the hypothesis that T1 is mostly comprised of commuters. T1 also has the lowest average interval of all the groups. This indicates that a higher percentage of passengers in T1 use the network more regularly than those in the other groups.

T2 almost has no peaks and has a modal interval of 12–13 h with an average interval of 33 h. Similarly, T4 does not have as distinct a peak as T1 or T3. It has a modal interval of 17–18 h and an average interval of 96 h. This indicates

Table 3 Summary of spatiotemporal trip pattern typology characteristics

Characteristic	T1	T2	T3	T4
Number of passengers	218	836	706	1441
Average hourly boardings	5	6	3	3
Average daily boardings	49	70	15	21
Average monthly boardings	1280	1750	260	490
Peak ridership	7a.m. & 4p.m	7a.m. & 3p.m	4p.m	3p.m

Table 4 Statistics of the passenger average time interval between activations (rounded up to the nearest hour)

Typology	Mean (hrs)	Median (hrs)	Modal interval (hrs)
T1	15	14	[9,10)
T2	33	28	[12,13)
T3	40	18	[3,4)
T4	96	64	[17,18)

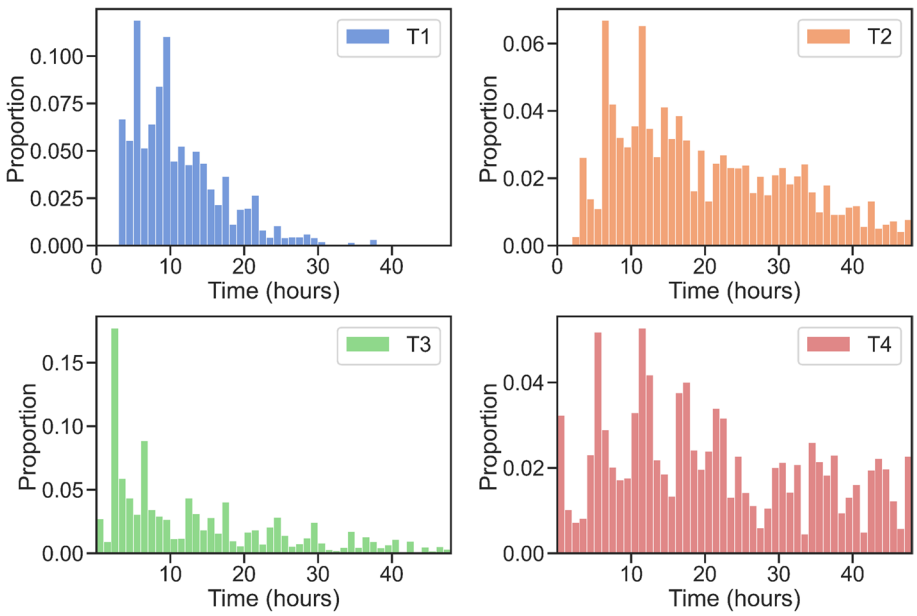

Fig. 6 Histograms of the passenger average time interval between activations

that many passengers in T2 and T4 might not use the bus on a regular schedule.

Distribution of Hourly Activations

We consider the hourly distribution of activations (rides) per typology, as shown in Fig. 7. The trends indicate a clear distinction in travel patterns between the four groups. Typologies T1 and T2 are characterized by a morning–afternoon travel pattern. T1 has two equal peaks of 40 activations at 7 a.m. and 4 p.m. T2 has a similar peak at 7 a.m. but a 50% larger peak of 60 activations at 3 p.m. The morning–afternoon pattern is most likely indicative of a daily commute. T3 and T4 also have two peaks during the day. The first peaks for T3 and T4, both with 10 activations, occur at 6 a.m. and 7 a.m. respectively. The second peak at 3 p.m. for both T3 and T4, with 10 and 20 activations respectively. T3 and T4 are different from the other typologies in that they have a much lower activation count of 10–16 activations (around 75% lower).

Passengers who belong to T2 and T4 also mostly travel during the evening. This could indicate that many passengers in these typologies largely use the bus for leisure or other activities, rather than for commuting regularly to a day job. We hypothesize that T2 and T4 may include a higher proportion of non-commuting passengers, such as the elderly, disabled, and children. Additionally, T2 and T4 may also include students. However, we note that there can be variations within the student population. Some students may have schedules that resemble regular commuting patterns, while others may engage in activities that result in non-commuting patterns.

Weekday and Weekend Distribution

An investigation of typology travel patterns by day of the week yields further interesting insights into the behavior of passengers. Figure 8 shows the weekly pattern of activations averaged by hour specific to a day of the week (a) and also weighted by the number of passengers in that hour (b). Weighting the hourly average activations by the number of passengers allows us to observe the relative frequency of passenger ridership in each typology by day of the week.

Figure 8(a) indicates trip volumes for each day of the week. The patterns correspond, as expected, to those in Fig. 7. We observe that the hourly average activations are significantly lower on the weekend in typologies T1 and T2, compared to the other two.

Observing the passenger–weighted average hourly activation distribution (Fig. 8(b)), we see that T1 has the highest average activations per user peaking at 0.24 activations per hour. It also has regular weekday morning–afternoon peaks. These observations further provide support for the characterization of this typology as being largely a commuter typology. Typologies T2 and T3 have a somewhat similar regularity in usage, with T2 having stable morning–afternoon patterns. However, T3 does not have uniform peaks throughout the week. The hourly average for both T2 and T3 is between 0.05 and 0.10 during weekdays. T4 has the least average activations per user with a maximum of 0.05 activations per hour, which means that T4 has the fewest regular passengers, with the peaks being consistently in the afternoon during weekdays. This further supports our previous analysis that members of this typology use the bus infrequently and for activities other than regular commuting.

On Saturdays, T1, T2, and T4 each have a single peak and significantly lower usage and regularity. However, T3 gains two equal morning–afternoon peaks and retains the same activation count but with lower regularity. On Sundays, all typologies have two peaks with even lower usage and regularity.

Seasonality Analysis

In this section, we analyze the seasonality and temporal patterns of the four typologies based on the activation data

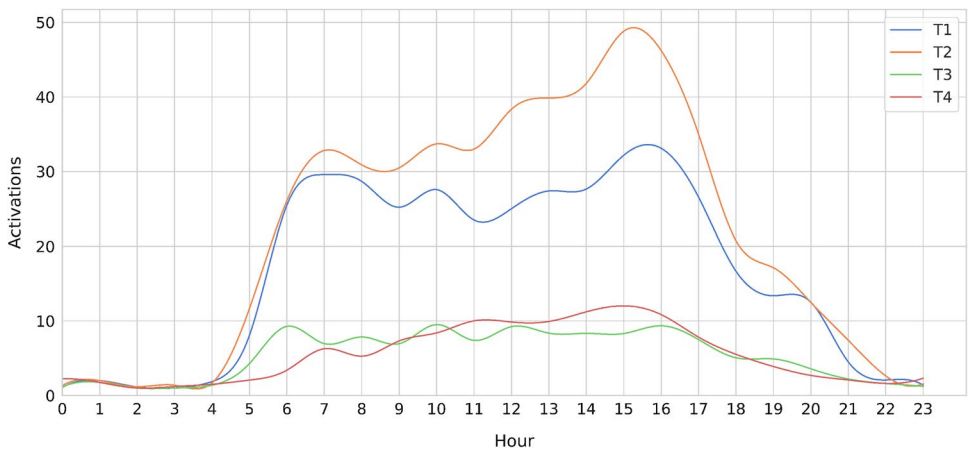

Fig. 7 Average hourly ridership by typology

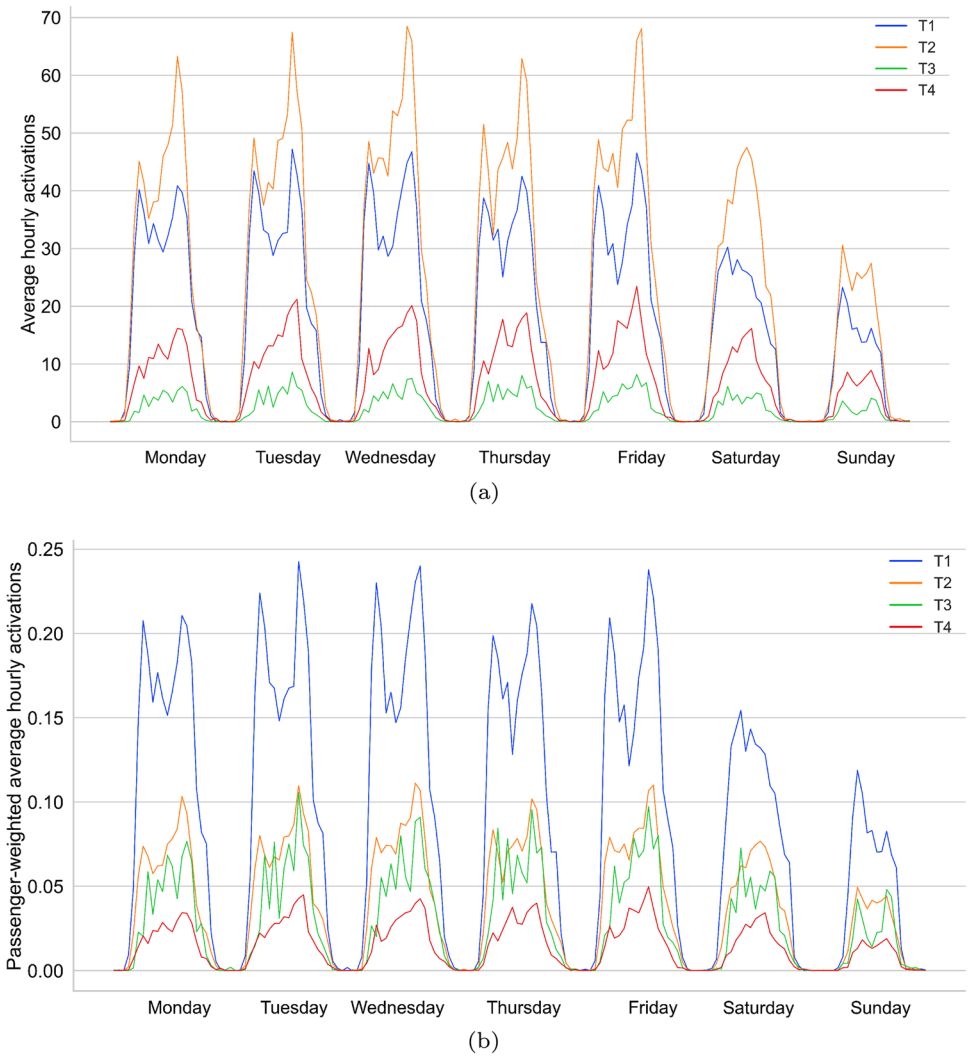

Fig. 8 Weekly activation patterns by typology: (**a**) hourly average and (**b**) passenger–weighted hourly average

from July 2020 to March 2021. The ridership seasonality results for each typology are presented in Fig. 9. Seasonal variations can be observed in the data, reflecting changes in ridership throughout the year.

In our analysis of ridership seasonality for the four typologies of passengers, we used an additive seasonal decomposition method (implemented via the Python module, statsmodels (Taylor 2010)). The time series of activations $z_k \in \mathbb{R}^T$ in each typology k is decomposed for each observation $z_{k,t}$ in time t as follows:

$$z_{k,t} = \tau_{k,t} + s_{k,t} + e_{k,t} \tag{6}$$

where $\tau_{k,t}$ is the trend component, $s_{k,t}$ the seasonal component, and $e_{k,t}$ the residual component at time t.

The observed data for T1 and T2 shows a pattern of initially increasing ridership followed by fluctuations around 600 activations. This pattern may indicate that both T1 and T2 represent regular commuters, such as workers and a subset of students who have schedules resembling regular commuting patterns (e.g. full-time students with fixed class schedules). The trend component confirms this observation, as both typologies exhibit dips before the start of December and January, possibly reflecting changes in work or school schedules during the holiday season.

T3, on the other hand, starts with very low activations and remains at that level until February, after which it rises sharply. The trend component further supports this observation, as it shows a steady increase in ridership beginning in February. This pattern could suggest that T3 represents a group of passengers whose ridership is influenced by factors that became prominent around February. Considering the timing and the sharp increase, T3 might represent a category of passengers such as seasonal workers or participants in educational programs and activities that start at this time of the year.

Lastly, T4 exhibits a pattern of initially moderate ridership that gradually declines over time. This behavior could

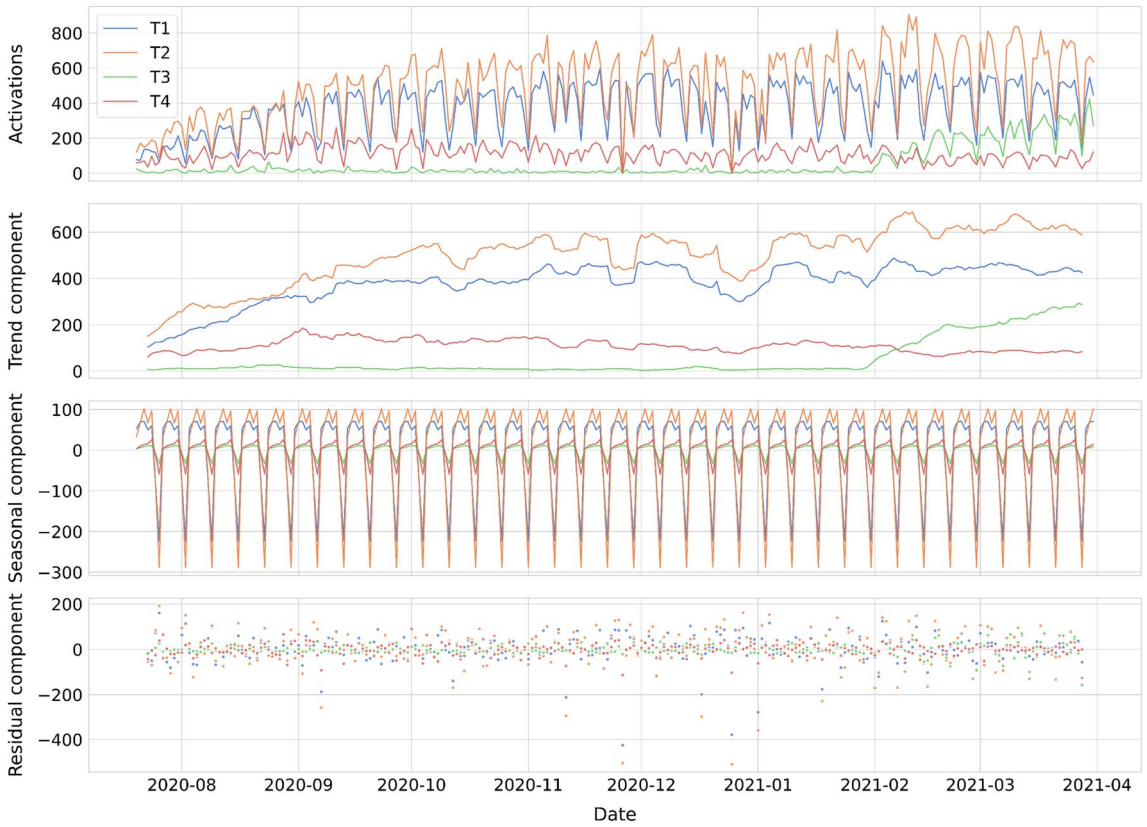

Fig. 9 Ridership seasonality decomposition by typology

indicate that T4 represents passengers who are not regular commuters, such as shoppers or other occasional travelers. The trend component shows a decline for T4, which may be a result of these passengers gradually adapting to alternative transportation options or reducing their travel frequency due to external factors.

The seasonal component demonstrates a weekly cycle for all typologies, highlighting the impact of weekly factors such as work or school schedules on each passenger group. On Sundays, activations decrease by around 65% in T1 and T2, by 60% in T3, and by 55% in T4. The higher dips in ridership for T1 and T2 on Sundays further support the inference that these typologies are likely associated with regular commuters such as workers, who have a more significant reduction in travel demand during weekends. In contrast, the lower dips in ridership for T3 and T4 on Sundays, indicate that these passenger groups potentially represent non-commuters such as shoppers or occasional travelers who have more varied travel patterns and are less influenced by the weekly work or school schedules.

The residual component indicates that most of the variation in the ridership data has been captured by the trend and seasonal components, as the residuals are centered around zero for all typologies. The average residuals are 0.236, −0.014, −0.502, and 0.008 for T1, T2, T3, and T4, respectively. However, T1 and T2 exhibit some high residuals between December and January, suggesting irregular fluctuations or potential outliers in the ridership data for these typologies during this period.

Spatial Patterns

To better comprehend the usage patterns of various passenger groups on the bus network and enhance the service planning abilities of transit agencies, we examine the spatial distribution and trends among different typologies. This involves studying the geographical distribution of boarding locations and the intensity of usage across different areas. We represent the spatial patterns by using heatmaps for each typology (see Fig. 10), which depict the concentration of boarding locations. The heatmaps have been generated using the Python library, Folium (Story 2013).

In T1, Springfield, the major urban center of the region, has the highest demand, with approximately 5600–5700 passengers boarding at various locations. Flows emanating from Springfield reveal demand extending to suburban areas, exemplified by Holyoke and Chicopee to the north, and Ludlow to the east, which register around 5300, 5600, and 5000

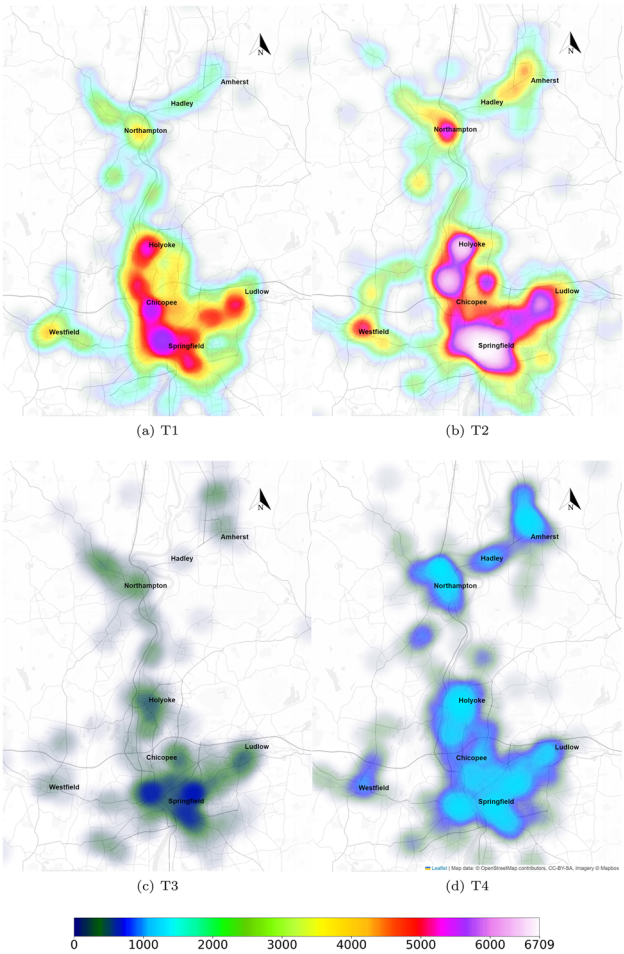

Fig. 10 Heatmaps depicting the spatial distribution of boardings for each typology from July 2020 through March 2021

boardings, respectively. This suggests a group that uses the bus service for extensive commuting within Springfield and neighboring areas.

T2 shows a similar pattern, with Springfield being the primary high-demand location. However, there is a noticeable upswing in demand in Northampton, Holyoke, and Chicopee, with around 5300, 6700, and 4700 boardings, respectively. Additionally, Amherst and Hadley, home to the flagship state university campus—University of Massachusetts Amherst (UMass), collectively account for a moderate demand of around 4100 boardings. This indicates that this group might be using the bus service for varied purposes, including commuting and academic activities.

In contrast, T3 exhibits a different pattern. The central part of Springfield, as well as East Springfield, remain the principal locations with around 700–800 boardings. The demand in other areas around Springfield, such as Holyoke and Chicopee, is significantly reduced compared to T1 and T2, suggesting more localized travel patterns within Springfield. Outside Springfield, the boarding figures dwindle further; Amherst, Hadley, and Westfield register a mere 200–300 boardings, respectively, implying limited intercity travels by this group of passengers.

Conversely, T4 presents a more heterogeneous and balanced demand across several areas. Springfield, Northampton, Amherst, Hadley, Holyoke, Chicopee, and Westfield all exhibit similar boarding figures of around 1200–1300. This equilibrium suggests a broad user base with diverse needs. The comparable levels of boarding across these areas indicate that passengers in T4 are likely engaging in various activities such as shopping, leisure, and accessing services.

Understanding these spatial patterns can be beneficial for transit agencies in resource allocation and service planning, as they reveal distinct trends among the typologies with regards to how different user groups utilize the bus network.

Faretype Analysis

Examining faretype usage frequency and their distribution among passengers is important for analyzing the passenger typologies. The PVTA network offers 12 faretypes, including *One Ride*, *Transfer*, *31-day Regular*, *College Pass*, *Children's One Ride*, and options for the elderly and disabled (*E&D*).

Frequency of Usage

Figure 11 displays the distribution of faretypes across the four typologies over an 8-month period.

We observe that the passengers in T2 dominate the usage of *31-day E&D* and *31-day Regular* faretypes, constituting approximately 54% and 39%, respectively. T1 also has a significant portion of the usage of these faretypes at around 17% and 21% respectively, which reinforces the inference of the regular commute pattern. Notably, no *Children's One Ride* faretypes are used in T1 and T2. Passengers in T2 account for 44% of the *Transfer* faretypes, followed by those in typologies T4 (33%) and T3 (18%). A negligible portion of passengers in T1 use the *Transfer* faretypes. However, at less than 2%, *Transfer* faretypes constitute a negligible percentage of the total faretypes used, which could mean that most people who need transfers regularly use a different type of faretype.

We also find that the *Day Pass* and the *One Ride* faretypes are predominantly used by passengers in T4, accounting for around 50% of both faretypes, followed by T2 (around 24%), T3 (around 19%), and T1 (around 7%). This suggests that T4 might be associated with passengers who are not regular commuters, possibly using the bus service for leisure or other activities.

Most bus rides are primarily associated with four faretypes. Specifically, 46.7% of the total activations are attributed to the *One Ride* faretype, 25.7% to the *Day Pass*, 8.5%

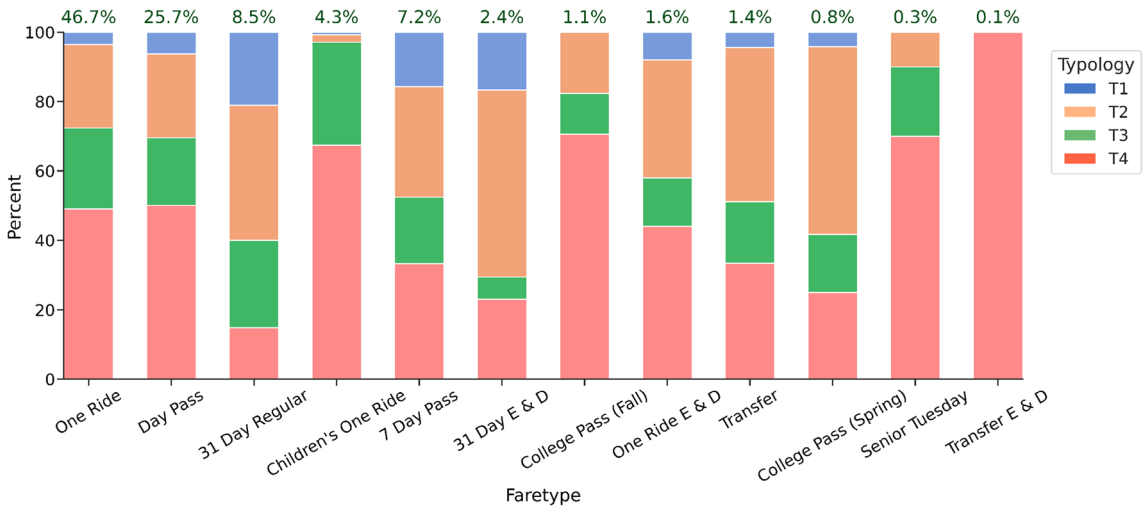

Fig. 11 Typology distribution in each faretype. Each bar is annotated with the faretype usage frequency across all activations

to the *31-day Regular*, and 7.2% to the *7-day Pass*. The *One Ride* faretype, being the most prevalent choice, suggests that a substantial number of bus rides are single-trip journeys. The notable preference for the *Day Pass*, constituting 25.7% of activations, indicates that a significant portion of bus rides are likely part of multi-destination trips within a single day. The *31-day Regular* and *7-day Pass*, accounting for 8.5% and 7.2% of activations respectively, are probably favored by regular commuters who rely on the bus network for their daily transportation needs over longer periods.

Further analyses of faretype usage across the typologies can provide insights to effectively target relevant groups to increase the usage of a particular faretype. The typology analysis can also facilitate the modification of existing faretypes or the design of new ones to potentially increase bus ridership in the network.

Temporal Faretype Patterns

We analyzed hourly usage by faretype (Fig. 12) to infer the activities of passengers with the most prevalent patterns in each typology. In T1, the *31-day Regular* faretype mostly used by commuters was the only faretype with a morning–afternoon peak pattern that coincides with the peaks we observed in Fig. 7. Other faretype patterns in T1 were negligible.

In T2, the *31-day Regular* faretype passengers had a morning peak that was not matched by an afternoon one. Passengers in this typology are thus likely to be commuters with a fixed morning shift. The second most frequently used faretype was the *31-day E&D* for elderly and disabled passengers, which peaks at 7 a.m. and continues at a steady demand level until 4 p.m. This indicates that T2 potentially comprises significant numbers of elderly or disabled passengers, along with those who do not use the bus to commute for full-day working hours.

In T3, the most frequently used faretype was the *31-day Regular* with morning–afternoon patterns. T3 is also the first to have any *Children's One Ride* faretypes used. It likely comprises regular commuters who have a fixed afternoon shift, as opposed to the fixed morning shift in T2, as well as many other passengers who use the bus less regularly but have similar spatial patterns.

Finally, there were two major peaks in T4 of 8 and 12 activations from passengers using the *One Ride* faretype at 10 a.m. and 2 p.m., respectively, and one peak of 7 activations at 12 p.m. from passengers who use the *Day Pass*. It also had a peak at 2 p.m. by *College Pass (Spring)* faretype users. This typology is characterized by non-uniform patterns and non-commuting parties.

Spatial Faretype Patterns

In order to analyze the spatial distribution of faretype usage, we plotted all trip trajectories within the study period to detect major patterns across the typologies (Fig. 13). In T1, the *7-day Pass* is mostly used outside eastern Springfield while the *Day Pass* faretype is mostly used inside eastern Springfield. In T2, the *One Ride* is the most frequently used faretype in eastern Springfield, while the *Day Pass* and *31-day Regular* faretypes are used for long travel. T3 has a similar pattern to T2 in that the *Day Pass* faretype is used mostly for long travel, while the *7-day Pass* and *One Ride* are the dominant faretypes in eastern Springfield. In T4, the most used faretypes are the *Day Pass* followed by the *One Ride* faretype with 93% of the total usage. They are also used less inside Springfield and more for longer trips.

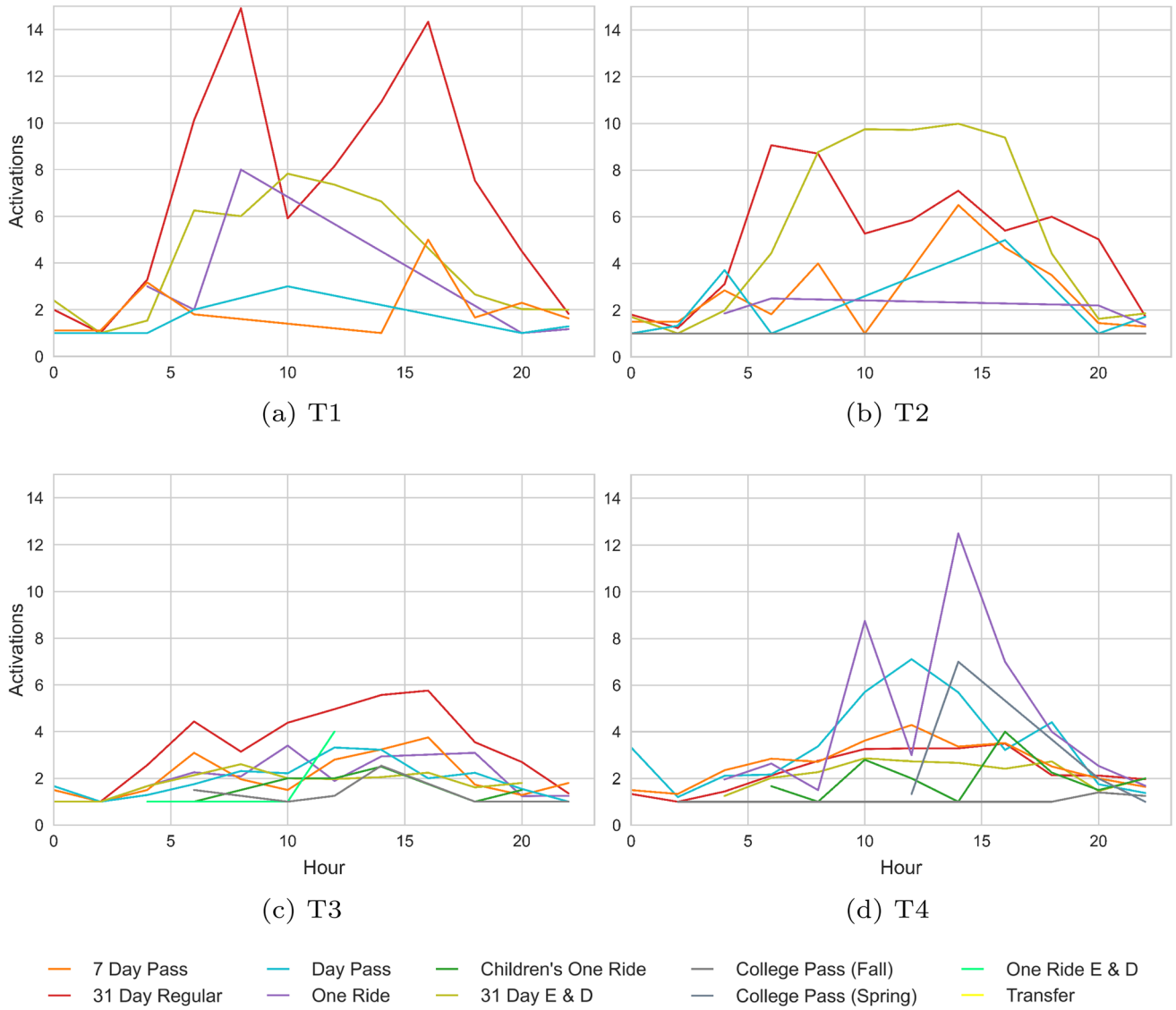

Fig. 12 Average hourly activations by typology and faretype

Summary

We have demonstrated the capability of detecting distinct trip pattern typologies using only noisy activation data from a small group of passengers of a regional bus network. To infer origins (boarding locations), we compared the performance of a spatial clustering approach (via DBSCAN) to that of a greedy assignment based on the nearest distance. The greedy approach performed consistently better across a range of time horizons and successfully inferred the boarding stop locations of 86% of the activations in the dataset. We then generated a dissimilarity matrix of pairwise separations between the time series of each passenger via the AWarp dynamic time warping algorithm for aligning sparse time series. We then conducted hierarchical agglomerative clustering using the Ward method to discover the prevailing trip patterns based on the activation time series of the passengers. This resulted in four spatiotemporal trip pattern typologies.

For T1, the patterns suggest that the majority might be regular commuters, based on consistent time activations and significant usage of the *31-day Regular* faretype. T2 appears to comprise a mix of individuals, including some who may be morning workers with unfixed work shifts, and possibly regular elderly or disabled travelers, as inferred from the variability in morning activation times and consistency in other time slots, as well as the dominance in the usage of the *31-day E&D* and *31-day Regular* faretypes. T3's characteristics seem to suggest that it may be composed of workers with a fixed shift start or end time in the afternoon, but no fixed

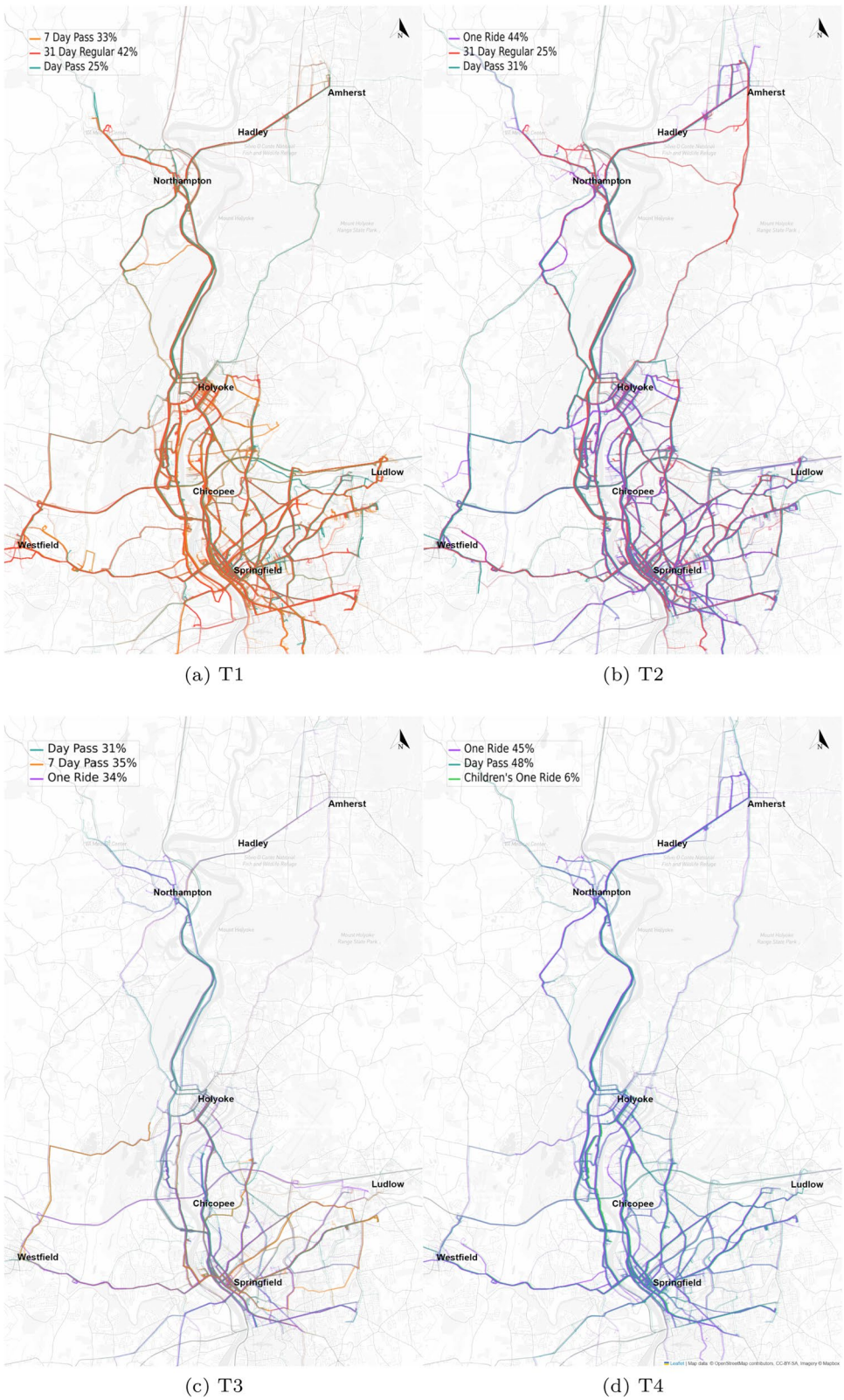

Fig. 13 Spatiotemporal trajectories by faretype used

shift in the morning. This is inferred from the consistency in the afternoon activation times, variability in the morning, and the usage of the *31-day Regular* faretype. T4 appears to be the most diverse group, possibly consisting of individuals such as students with irregular schedules and shoppers, which is inferred from the lack of regular daily patterns in

activation times, and the predominant usage of the *Day Pass* and *One Ride* faretypes.

We note that the characterizations of the groups T1, T2, T3, and T4 are tentative and based on the patterns observed in the activation times and faretypes. We cannot definitively conclude the exact composition of these groups. The inferences are made based on patterns that emerge from the data and further research with socio-demographic data might be needed for a more conclusive characterization.

Conclusion

In this paper, we used mobile ticketing data to analyze spatiotemporal travel patterns in a regional bus network. Our case study was the Pioneer Valley Transit Authority bus network in the US state of Massachusetts. The data consisted of timestamped location-based activations collected via a mobile application—a class of automated fare collection systems—which are required prior to or slightly after boarding a bus on this network. We addressed two areas of inquiry: passenger origin inference and trip pattern analysis. Our method yielded four spatiotemporal trip pattern typologies.

Our approach to analyzing mobility patterns gave distinct results in the absence of high-cost surveys and provided valuable insights into travel behavior across the network, which can inform decision-making by urban planners and policymakers in the Pioneer Valley region. These insights may help to identify areas in need of service improvements or adjustments, leading to more efficient resource allocation and better overall service quality. By relying on data-driven analysis rather than traditional assumptions, this research demonstrates the potential for more informed and targeted decision-making in public transportation planning.

Given the diverse profiles and characteristics of transit systems across the world, the analyses of travel patterns, either on systems yet to be studied or via novel methods in prior study areas, will continue to yield important insights. In particular, the introduction and adoption of new modes of transportation (such as micromobility), mobility services, and technologies will likely continue to disrupt travel patterns and behavior, as well as produce larger datasets. These trends motivate the need for continued research and innovation in this area. Ultimately, we expect that our findings can serve as a basis for further research on trip patterns in bus networks with sparse mobile ticketing information. Given the technological and budgetary constraints in implementing automated fare collection systems, greater efficiency and parsimony will be required to provide planning insights for effective management and targeted service improvements, which in turn should serve to mitigate the decline in public transportation ridership.

Author Contributions The authors confirm their contribution to the paper as follows: study conception and design: JO, MM; data collection: MM; analysis and interpretation of results: MM, JO; draft manuscript preparation: MM, JO. Both authors reviewed the results and approved the final version of the manuscript.

Funding The research leading to these results received funding from the Federal Transit Administration under Grant Agreement ID FAIN MA-2021-012-00.

Data Availability The data used in the analyses were provided by the Pioneer Valley Transit Authority (PVTA). The data used in this study are not publicly available due to privacy concerns as the data contain sensitive personal information that cannot be shared without the explicit consent of the individuals involved. The code and documentation for generating the results presented in this paper are available via this GitHub repository.

Declarations

Conflict of Interest The authors declare that they have no competing interests.

Ethical Approval This study did not require ethical approval as it did not involve any human or animal subjects.

References

Agard B, Nia VP, Trépanier M (2013) Assessing public transport travel behaviour from smart card data with advanced data mining techniques. In: World Conference on Transport Research 13:13

Ait-Ali A, Eliasson J (2019) Dynamic origin-destination estimation using smart card data: an entropy maximisation approach. arXiv e-prints

Alsger AA, Mesbah M, Ferreira L et al (2015) Use of smart card fare data to estimate public transport origin-destination matrix. Transp Res Record 2535(1):88–96. https://doi.org/10.3141/2535-10

Asadi R, Regan A (2019) Spatio-temporal clustering of traffic data with deep embedded clustering. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Prediction of Human Mobility—PredictGIS'19. ACM Press, Chicago, pp 45–52, https://doi.org/10.1145/3356995.3364537

Asif MT, Dauwels J, Goh CY et al (2014) Spatiotemporal patterns in large-scale traffic speed prediction. IEEE Trans Intell Transp Syst 15(2):794–804. https://doi.org/10.1109/TITS.2013.2290285

Ben-Akiva ME, Morikawa T (1989) Data fusion methods and their applications to origin-destination trip tables. In: Transport Policy, Management & Technology towards 2001: Selected Proceedings of the Fifth World Conference on Transport Research, pp 279–293

Briand AS, Côme E, Trépanier M et al (2017) Analyzing year-to-year changes in public transport passenger behaviour using smart card data. Transp Res Part C: Emerg Technol 79:274–289. https://doi.org/10.1016/j.trc.2017.03.021

Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat 3(1):1–27. https://doi.org/10.1080/03610927408827101

Chen R, Zhang J, Ravishanker N et al (2019) Clustering activity-travel behavior time series using topological data analysis. J Big Data Anal Transp 1(2):109–121. https://doi.org/10.1007/s42421-019-00008-6

Chen E, Ye Z, Wang C et al (2020) Subway passenger flow prediction for special events using smart card data. IEEE Trans Intell

Transp Syst 21(3):1109–1120. https://doi.org/10.1109/TITS.2019.2902405
Costa MA, Marra AD, Corman F (2023) Public Transport Commuting Analytics: A Longitudinal Study Based on GPS Tracking and Unsupervised Learning. Data Sci Trans 5(3). https://doi.org/10.1007/s42421-023-00077-810.1007/s42421-023-00077-8
Cournapeau D (2007) Scikit-learn: machine learning in Python. https://scikit-learn.org/stable/
Cui A (2006) Bus passenger origin-destination matrix estimation using automated data collection systems. Thesis, Massachusetts Institute of Technology
Decouvelaere R, Trépanier M, Agard B (2022) Modulated spatiotemporal clustering of smart card users. Public Transport. https://doi.org/10.1007/s12469-022-00305-4
El Mahrsi MK, Côme E, Oukhellou L et al (2017) Clustering smart card data for urban mobility analysis. IEEE Trans Intell Transp Syst 18(3):712–728. https://doi.org/10.1109/TITS.2016.2600515
Ester M, Kriegel HP, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd 96(34):226–231
Ge Q, Fukuda D (2016) Updating origin–destination matrices with aggregated data of GPS traces. Transp Res Part C: Emerg Technol 69:291–312. https://doi.org/10.1016/j.trc.2016.06.002
Giannotti F, Nanni M, Pinelli F, et al (2007) Trajectory pattern mining. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, KDD '07, pp 330–339, https://doi.org/10.1145/1281192.1281230
Gordon JB (2012) Intermodal passenger flows on London's public transport network: automated inference of full passenger journeys using fare-transaction and vehicle-location data. Thesis, Massachusetts Institute of Technology
Hanson S, Huff J (1986) Classification issues in the analysis of complex travel behavior. Transportation 13(3):271–293. https://doi.org/10.1007/BF00148620
Hanson S, Huff OJ (1988) Systematic variability in repetitious travel. Transportation 15(1):111–135. https://doi.org/10.1007/BF00167983
Hazelton ML (2010) Statistical inference for transit system origin-destination matrices. Technometrics 52(2):221–230. https://doi.org/10.1198/TECH.2010.09021
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. Transp A: Transport Sci 16(1):56–75. https://doi.org/10.1080/23249935.2018.1479722
Hochmair HH (2016) Spatiotemporal pattern analysis of taxi trips in New York City. Transp Res Record 2542(1):45–56. https://doi.org/10.3141/2542-06
Hwang Y, Gelfand SB (2018) Constrained sparse dynamic time warping. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp 216–222, https://doi.org/10.1109/ICMLA.2018.00039
Inmook LEE (2019) Estimating of bus-trip destinations using temporal travel patterns of smart card data. Thesis, Seoul National University
Jones P, Clarke M (1988) The significance and measurement of variability in travel behaviour. Transportation 15(1):65–87. https://doi.org/10.1007/BF00167981
Kahana D, Dickens M (2023) APTA POLICY BRIEF Transit Ridership. Tech. rep, APTA
Kisilevich S, Mansmann F, Nanni M et al (2010) Spatio-temporal clustering. In: Maimon O, Rokach L (eds) Data mining and knowledge discovery handbook. Springer, US, Boston, pp 855–874. https://doi.org/10.1007/978-0-387-09823-4_44
Liu L, Miller HJ, Scheff J (2020) The impacts of COVID-19 pandemic on public transit demand in the United States. PLoS One 15(11):e0242476. https://doi.org/10.1371/journal.pone.0242476
Liu X, Van Hentenryck P, Zhao X (2021) Optimization models for estimating transit network origin–destination flows with big transit data. J Big Data Anal Trans 3(3):247–262. https://doi.org/10.1007/s42421-021-00050-3
Ma X, Wu YJ, Wang Y et al (2013) Mining smart card data for transit riders' travel patterns. Transp Res Part C: Emerg Technol 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
Mohammed M, Oke J (2023) Origin-destination inference in public transportation systems: A comprehensive review. Int J Trans Sci Technol 12(1):315–328. https://doi.org/10.1016/j.ijtst.2022.03.002
Mueen A, Chavoshi N, Abu-El-Rub N et al (2018) Speeding up dynamic time warping distance for sparse time series data. Knowl Inform Syst 54(1):237–263. https://doi.org/10.1007/s10115-017-1119-0
Navick D, Furth P (1994) Distance-based model for estimating a bus route origin-destination matrix. Transportation research record, p 16
Nishiuchi H, King J, Todoroki T (2013) Spatial-temporal daily frequent trip pattern of public transport passengers using smart card data. Int J Intell Transp Syst Res 11(1):1–10. https://doi.org/10.1007/s13177-012-0051-7
O'Toole R (2018) Charting public transit's decline. https://www.cato.org/policy-analysis/charting-public-transits-decline
Pas EI (1987) Intrapersonal variability and model goodness-of-fit. Transp Res Part A: Gen. https://doi.org/10.1016/0191-2607(87)90032-X
Pas EI, Koppelman FS (1986) An examination of the determinants of day-to-day variability in individuals' urban travel behavior. Transportation 13(2):183–200. https://doi.org/10.1007/BF00165547
Prasannakumar V, Vijith H, Charutha R et al (2011) Spatio-temporal clustering of road accidents: GIS based analysis and assessment. Procedia Soc Behav Sci 21:317–325. https://doi.org/10.1016/j.sbspro.2011.07.020
Primerano F, Taylor MAP, Pitaksringkarn L et al (2008) Defining and understanding trip chaining behaviour. Transportation 35(1):55–72. https://doi.org/10.1007/s11116-007-9134-8
PVTA (2023) About PVTA. http://www.pvta.com/about.php
Rinzivillo S, Pedreschi D, Nanni M et al (2008) Visually-driven analysis of movement data by progressive clustering. Inform Vis 7:225–239. https://doi.org/10.1057/palgrave.ivs.9500183
Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7
Sakoe H, Chiba S (1978) Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans Acoust Speech Signal Process 26(1):43–49. https://doi.org/10.1109/TASSP.1978.1163055
Salvador S, Chan P (2007) Toward accurate dynamic time warping in linear time and space. Intell Data Anal 11(5):561–580. https://doi.org/10.3233/IDA-2007-11508
Sanaullah I, Alsaleh N, Djavadian S et al (2021) Spatio-temporal analysis of on-demand transit: a case study of Belleville, Canada. Transp Res Part A: Policy Pract 145:284–301. https://doi.org/10.1016/j.tra.2021.01.020
Shao F, Sui Y, Yu X et al (2019) Spatio-temporal travel patterns of elderly people—a comparative study based on buses usage in Qingdao, China. J Transport Geogr 76:178–190. https://doi.org/10.1016/j.jtrangeo.2019.04.001

Shen D, Chi M (2021) TC-DTW: accelerating multivariate dynamic time warping through triangle inequality and point clustering. https://doi.org/10.48550/arXiv.2101.07731

Shi Z, Pun-Cheng LSC (2019) Spatiotemporal data clustering: a survey of methods. ISPRS Int J Geo-Inform 8(3):112. https://doi.org/10.3390/ijgi8030112

Shi Z, Pun-Cheng LSC, Liu X et al (2020) Analysis of the temporal characteristics of the elderly traveling by bus using smart card data. ISPRS Int J Geo-Inform 9(12):751. https://doi.org/10.3390/ijgi9120751

Shokoohi-Yekta M, Hu B, Jin H et al (2017) Generalizing DTW to the multi-dimensional case requires an adaptive approach. Data Min Knowl Discov 31(1):1–31. https://doi.org/10.1007/s10618-016-0455-0

Song J, Zhao C, Zhong S et al (2019) Mapping spatio-temporal patterns and detecting the factors of traffic congestion with multi-source data fusion and mining techniques. Comput Environ Urban Syst 77(101):364. https://doi.org/10.1016/j.compenvurbsys.2019.101364

Story R (2013) Folium. https://python-visualization.github.io/folium/

Strauss T, von Maltitz MJ (2017) Generalising ward's method for use with Manhattan distances. PLoS One 12(1):e0168288. https://doi.org/10.1371/journal.pone.0168288

Sun Y, Xu R (2012) Rail transit travel time reliability and estimation of passenger route choice behavior: analysis using automatic fare collection data. Transp Res Record 2275(1):58–67. https://doi.org/10.3141/2275-07

Sun D, Zhang K, Shen S (2018) Analyzing spatiotemporal traffic line source emissions based on massive didi online car-hailing service data. Transp Res Part D: Transport Environ 62:699–714. https://doi.org/10.1016/j.trd.2018.04.024

Taylor J (2010) Statsmodels: statistical modeling and econometrics in Python. https://www.statsmodels.org/stable/index.html

Wang W (2010) Bus passenger origin-destination estimation and travel behavior using automated data collection systems in London. Thesis, Massachusetts Institute of Technology, UK

Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236–244. https://doi.org/10.1080/01621459.1963.10500845

Yong J, Zheng L, Mao X et al (2021) Mining metro commuting mobility patterns using massive smart card data. Phys A: Stat Mech Appl 584(126):351. https://doi.org/10.1016/j.physa.2021.126351

Zhang F (2022) Not all extreme weather events are equal: impacts on risk perception and adaptation in public transit agencies. Clim Change 171(1):3. https://doi.org/10.1007/s10584-022-03323-0

Zhang F, Welch EW, Miao Q (2018) Public organization adaptation to extreme events: mediating role of risk perception. J Public Admin Res Theory 28(3):371–387. https://doi.org/10.1093/jopart/muy004

Zhao J, Tian C, Zhang F, et al (2014) Understanding temporal and spatial travel patterns of individual passengers by mining smart card data. In: 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp 2991–2997, https://doi.org/10.1109/ITSC.2014.6958170

Zhao J, Qu Q, Zhang F et al (2017) Spatio-temporal analysis of passenger travel patterns in massive smart card data. IEEE Trans Intell Transp Syst 18(11):3135–3146. https://doi.org/10.1109/TITS.2017.2679179

Zhong S, Sun DJ (2022) Analyzing spatiotemporal congestion pattern on urban roads based on taxi GPS data. In: Zhong S, Sun DJ (eds) Logic-driven traffic big data analytics: methodology and applications for planning. Springer Nature, Singapore, pp 97–118. https://doi.org/10.1007/978-981-16-8016-8_5

Zhong C, Manley E, Arisona SM et al (2015) Measuring variability of mobility patterns from multiday smart-card data. J Comput Sci 9:125–130. https://doi.org/10.1016/j.jocs.2015.04.021

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com