Data Science for Transportation (2024) 6:19
https://doi.org/10.1007/s42421-024-00108-y

RESEARCH

Enhanced Seasonal Typology-Informed Transit Trip Chaining via Mobile Boarding and Survey Data

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

Received: 13 February 2024 / Revised: 22 July 2024 / Accepted: 13 August 2024

Abstract
Trip chaining is essential in estimating transit demand models based on big data sources, which hold advantages over traditional data collection methods in terms of both accuracy and cost-effectiveness. However, regional transit systems, accounting for 60% of transit systems in the United States, are often not equipped with smart card systems that are typically needed for trip chaining due to the high costs involved. This study introduces an innovative trip chaining framework that leverages boarding-only mobile ticketing automated fare collection (AFC) data. Our approach incorporates spatiotemporal passenger types obtained via typology analysis. This is followed by calibrating type-specific and season-aware parameters using survey data supplied by the network operators. We then trained a gradient boosting machine to learn the error structure of the initial trip chaining flow estimates and augment final trip frequency predictions. Our findings demonstrate that incorporating spatial error correction significantly improves the trip chaining results. However, the best performance is achieved when season-awareness and typology information are integrated with spatial error correction, reducing the mean absolute error (MAE) by approximately 70% and the symmetric mean absolute percentage error (SMAPE) by 85%. Furthermore, the typology, which we validate with a classification model, provides a deeper understanding of passenger travel characteristics which can facilitate further planning efforts. The spatial error correction model also provides insights into portions of the network that can be better targeted for survey data collection and mobile ticketing adoption to improve future models. Overall, this research will enable regional transit planners to efficiently utilize limited resources and generate valuable insights that would otherwise remain inaccessible.

Keywords Trip chaining · Typology analysis · Machine learning · Public transit · Automated fare collection · Mobile ticketing

Introduction

Accurately capturing and understanding passenger travel patterns is a significant challenge in public bus transit systems. Traditional methods of data collection have often been resource-intensive and not always feasible (Vanderwaart 2016; Mohammed and Oke 2023), leading to data and knowledge gaps. These limitations hamper the ability to make informed decisions and optimize resource allocation.

✉ 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

Recent advancements, however, have seen the emergence of automated data collection systems (ADCS), notably automated fare collection (AFC) smartcard-based closed systems. These offer high-frequency and high-resolution data, enabling the creation of individual full-trip records. Such detailed data are pivotal for implementing trip chaining, a method that infers complete passenger journeys, which can significantly enhance the accuracy of demand estimation models (Thill and Thomas 1987; Mohammed and Oke 2023). However, while smartcard-based AFC systems cover a significant portion of transit networks in the United States, smaller bus transit systems often struggle with resource constraints and rely on more accessible AFC systems, such as mobile ticketing systems (Goldstein 2016). However, there is a noticeable lack of trip chaining efforts in these open systems, presenting a significant challenge in accurately modeling passenger travel patterns in regional networks.

To address this gap, we propose a spatiotemporal trip chaining framework that harnesses the potential of boarding-only data from mobile ticketing systems within regional transit networks. The usage of mobile ticketing in open AFC systems, which typically collect boarding-only information, introduces complexities in accurately capturing passenger alighting data. This limitation is especially challenging in the context of mobile ticketing due to the lack of precise boarding information in addition to the complete lack of alighting information. This is compounded by the optional nature of mobile ticketing, which not only limits the sample size but also introduces biases based on smartphone ownership and user preferences toward mobile payments. Consequently, advanced data processing techniques are required to infer trip origins in addition to destinations.

We present a trip chaining framework that seeks to mitigate these challenges by integrating three key elements into the trip chaining framework: spatiotemporal passenger typology, seasonality, and spatial error correction. We derived the spatiotemporal passenger typology by clustering passenger trajectories to capture diverse travel patterns, providing type-specific parameters for the trip chaining process. Seasonality adjustments account for temporal variations in travel behavior, optimizing model parameters for different seasons. Finally, we trained a gradient boosting machine (GBM) to learn and correct spatial errors in trip chaining predictions. We used survey data to calibrate and validate the model at the zonal level. We show that our approach significantly reduces the test mean absolute error (MAE) by approximately 70% and the symmetric mean absolute percentage error (SMAPE) by 85% when compared to the survey data. We expect that our approach can enable regional transit planners to obtain insights from boarding-only data and, consequently, formulate strategic decisions that would have been beyond reach using the conventional approaches.

The rest of the paper is organized as follows. In Sect. 2, we provide an overview of the trip chaining in general and available data sources. We describe the data sources available to us and our trip chaining framework in Sect. 3. We then present and discuss our framework results in Sect. 4, and Sect. 5 provides our final remarks. We summarize the acronyms used throughout this paper in Table 6.

Background

Trip chaining, a concept that emerged in the 1960s within economic and transportation literature, originally analyzed passenger behavior and travel patterns using trip survey data (Thill and Thomas 1987). Often, research in transportation behavior predicates the rationality of decision-making and perceives trip chaining through the lens of utility maximization (Marble 1960). Consequently, early trip chaining endeavors were formulated in tandem with central place theory, which explains the spatial distribution of services and settlements based on their importance and accessibility (Thill and Thomas 1987).

As the complexity of factors, such as location, duration, and sequence of activities, became apparent, scholars began to break down travel into individual segments for detailed analysis (Lerman 1979; Horowitz 1980; Van Der Hoorn 1983; Kitamura 1984; Thill and Thomas 1987; O'Kelly 2010). Furthermore, since the late 1990s, a growing number of studies have used travel surveys to explore the intricacies of passenger travel choices. These studies also examined the impact of environmental factors on trip chains and investigated correlations with individual characteristics (Lee et al. 2002; Yang et al. 2007; Noland and Thomas 2007; Primerano et al. 2008; Schmöcker et al. 2010; Currie and Delbosc 2011; Yalamanchili et al. 1999).

In recent years, the advent of automated data collection systems (ADCS) has enabled public transit agencies to collect high-resolution and up-to-date data sets. The abundance of these data has spurred researchers to develop a variety of origin–destination estimation approaches employing trip chaining to be used in applications, such as transit planning and demand forecasting (Assemi et al. 2020). At its core, trip chaining connects passenger activities via travel throughout an entire journey (Primerano et al. 2008). With the emergence of new research avenues and increased data availability, trip chaining has become instrumental in constructing seed matrices, which are essential components of origin–destination (OD) models for transit.

This significant change in trip chaining has led to advanced techniques for creating comprehensive origin–destination matrices, known as *origin–destination–transfer* (ODX) models (Vanderwaart 2016; Sánchez-Martínez 2017). These models, supported by ADCS, which streamline the process of data collection and leads to more frequent and up-to-date ODX matrices, are increasingly surpassing the traditional OD models in terms of accuracy and efficiency (Gordon 2012; Cui 2006). The key to this progress is incorporating transfer inference within the trip chaining process, which links various trip segments to create a complete journey for passengers (Assemi et al. 2020; Alsger et al. 2016).

Historically, trip chaining predominantly found application in networks employing smart card AFC systems, which are invaluable in estimating travel demand and assisting decision-making processes. Smart cards offer an unparalleled level of data accuracy and tend to be widely adopted by passengers in the networks that support them. Moreover, in large cities, the sheer volume of data accessible through smart cards enhances the precision of models constructed using trip chaining methodologies. Consequently, there has been a prolific expansion of trip chaining approaches and refinements tailored to the large datasets collected through smart cards in urban settings (Mohammed and Oke 2023; Cui 2006; Gordon et al. 2013; Vanderwaart et al. 2017).

However, regional transit systems frequently employ more cost-effective non-smart card open AFC systems, which, in contrast to their smart card counterparts, offer data that are less detailed and accurate. Despite this, there remains a significant yet largely unexplored opportunity for trip chaining frameworks that can effectively utilize data from these non-smart card systems. While some studies have begun to explore trip chaining in these systems, this area remains relatively underdeveloped. Furthermore, existing research has not fully explored the use of typology via clustering in enhancing trip chaining methods. Typology-based approaches have the potential to improve trip chaining results while providing additional insights into characteristic passenger travel patterns. Additionally, while seasonal variations can significantly impact travel behavior (Liu et al. 2016), season-aware trip chaining model development is absent in the literature.

Our paper addresses these gaps by presenting a novel framework that adapts trip chaining for non-smart card open AFC systems and incorporates a typology-based approach. It also introduces a season-aware dimension to trip chaining modeling. By introducing these contributions, this paper offers new insights and methodologies that can greatly enhance our understanding and modeling of passenger travel behavior in various transit system settings.

Methodology

In this section, we outline our approach to analyzing travel patterns from boarding-only mobile ticketing data and performing trip chaining incorporating typology and seasonality considerations. We first created spatial trajectories for each user and clustered them to determine distinct behavioral patterns and to obtain user types. Subsequently, we developed a trip chaining framework to infer boarding, alighting, and transfer points. Next, we analyzed the seasonal patterns of each user type to adapt the framework's parameters to seasonal changes. Finally, we integrated a gradient boosting machine (GBM) to refine the framework's accuracy as compared with survey data by targeting unexplained data variations. The overall steps of our research framework are depicted in Fig. 1, and essential terms are summarized in Table 7.

Data

Study Area

We conducted our study on the Pioneer Valley Transit Authority (PVTA) bus network in Western Massachusetts, which serves 24 communities with 36 routes and 186 buses and connects 1,811 bus stops (Pioneer Valley Transit Authority 2023). A map of bus stops and boarding locations is shown in Fig. 2a.

Travel Survey Data

Travel surveys from 2019 and 2022 offer passenger start and end locations and routes used, including transfers. Figure 2b provides a visual representation of the survey data, showing the general locations of passenger boardings and alightings.

The survey data provide only general location information, indicating the broader area where users board buses, without specifying the exact stop identification number (ID). To address this, we used Fuzzy matching (Zadeh 1965) and

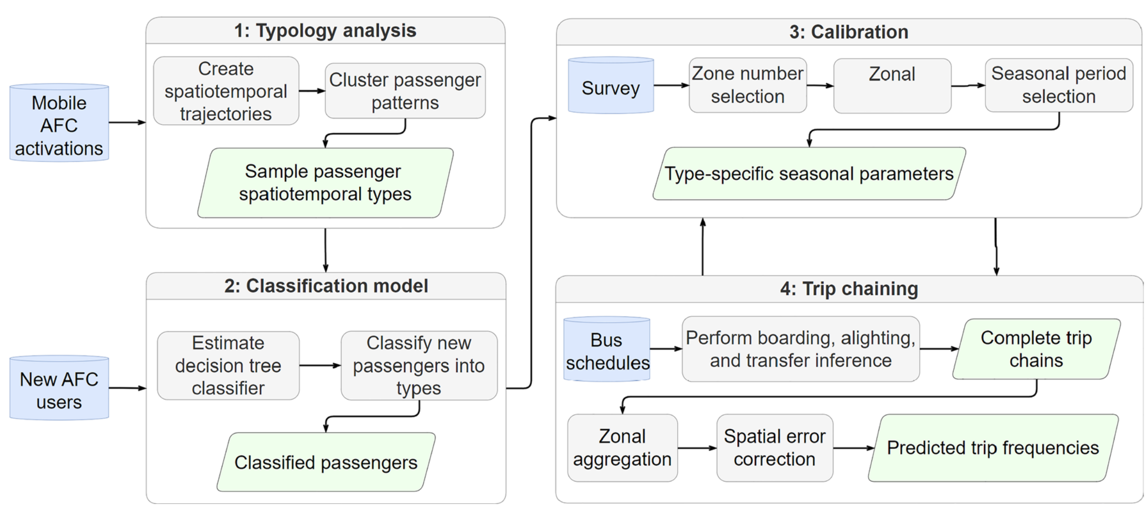

Fig. 1 Flowchart of the research framework

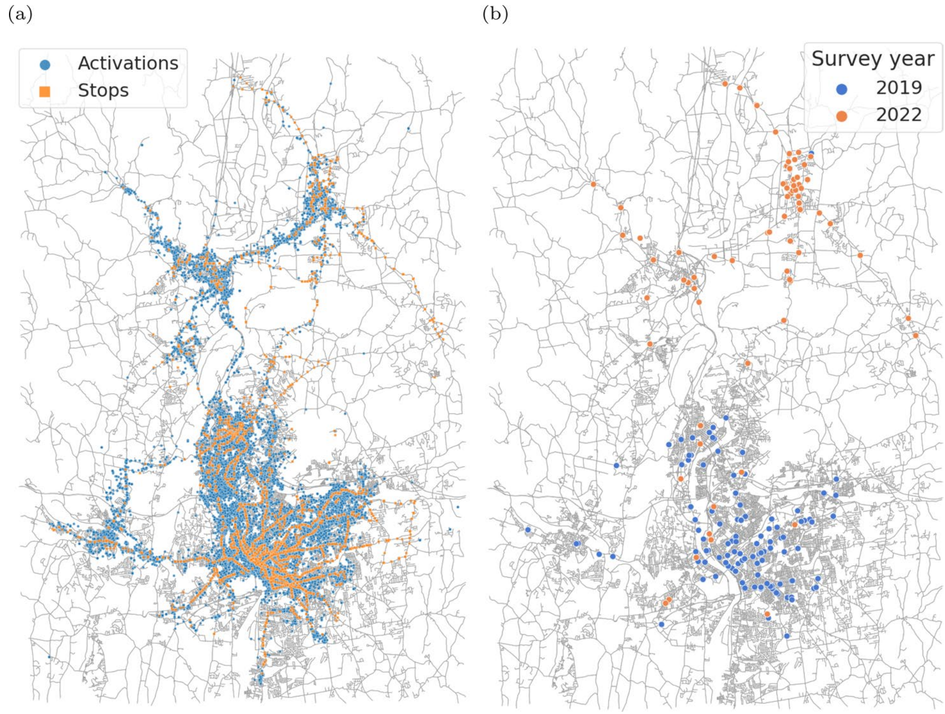

Fig. 2 PVTA network and travel survey data: **a** map of bus stops and the distribution of user activations in the PVTA network from July 2020 through December 2021, and **b** map of approximate passenger boarding locations collected from the survey data collected by the PVTA during the survey periods of 2019 and 2022

Levenshtein distance (Levenshtein 1966) to align survey responses with specific bus stops. We used Fuzzy matching to broadly align the general location names from the survey with our bus stop database, while the Levenshtein distance quantified the textual similarity between these names, refining the alignment process. Together, these methods successfully mapped 91% of boarding and 89% of alighting locations. For trips requiring transfers, we identified common stops between routes, with the median stop designated as the transfer location for 72% of cases.

However, this process introduces potential sources of error due to the inherent uncertainties in matching general location names with specific bus stops. These errors can impact the accuracy of the trip chaining framework, especially in areas with a high density of bus stops where similar names could lead to incorrect assignments. To optimize the dataset for analysis and minimize potential errors, we used the K-means clustering algorithm to group network stops into Z zones. These zones represented boarding, alighting, and transfer locations in our analysis. We then aggregated survey results at the zone level.

Automated Fare Collection Data

Unlike smart cards, which often collect detailed and accurate data about specific buses, trip IDs, routes, and directions, mobile ticketing systems primarily capture GPS locations at ticket activation without any direct connection to a specific bus. This necessitates the inference of boarding locations, which is typically not required with smart card-based data.

Since July 2020, the study network (PVTA) has been using a mobile app for fare collection, recording user identification numbers (IDs), timestamps, and GPS locations at ticket activation. The noisy data set, covering up to December 2021, presents challenges due to manual ticket validation by bus drivers and potential multiple or even false activations by passengers. For example, the GPS data can be inaccurate due to signal loss, and manual validation may introduce human errors. Furthermore, some activations may occur far from any bus stop or outside of operational hours, making it difficult to match activations with actual trips. Additionally, the dataset only partially reflects passenger boardings as the mobile ticketing system was adopted by a growing number of bus network riders, with less than 16% using it by the end of 2021.

While all mobile application users are included in the model, we discard approximately 20% of their activations, such as those with missing GPS data, indistinguishable multiple activations, and those not corresponding to valid trips (e.g., activations outside of operational hours). Analysis of the remaining mobile activations, comparing characteristics such as time of day, spatial distribution, and frequency of use to the original unfiltered set, confirmed their representativeness of the overall mobile ticketing population, as they maintained similar distributions in these key travel behavior aspects. The trends in weekly network usage and weekly new user count, fare type usage across the network, and the distribution of activations per user are visualized in Fig. 3a–c.

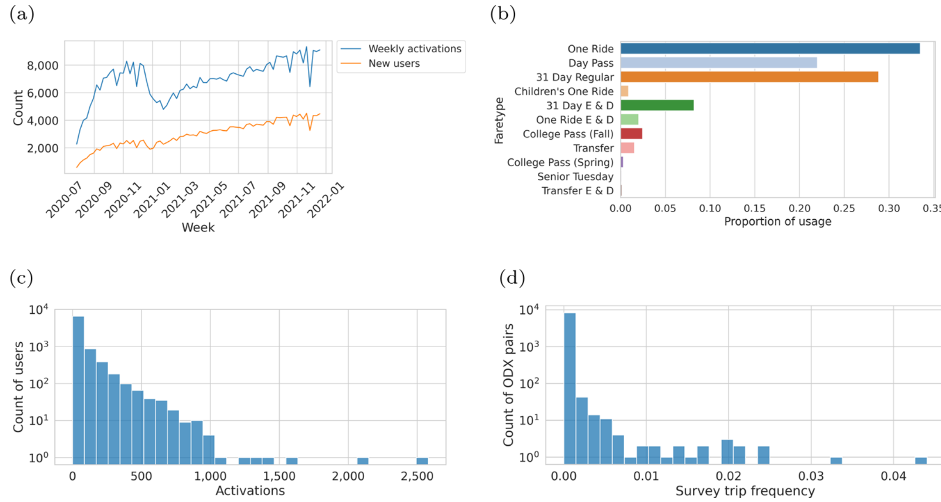

Fig. 3 Overview of PVTA mobile ticketing and survey data: **a** weekly ticket activations and unique active users in the PVTA network showing the rising demand of mobile ticketing, **b** faretype usage by all users throughout the study period from July 2020 to December 2021, **c** Histogram of ticket activations per user, and **d** Histogram of observed survey trip frequencies by zone

Trip Frequencies

Both the survey data and the trip chaining results are formatted as origin–destination–transfer seed matrices, with raw trip counts (x_{ijk}) for each unique combination of origin i, destination j, and transfer zones k ($k = 0$ signifies no transfers). To obtain trip frequencies (p_{ijk}), which represent the likelihood of trips between zones (the aggregation of stops into zones is discussed later in Sect. 3.5.1), we normalized these counts as follows:

$$p_{ijk} = \frac{x_{ijk}}{\sum\limits_{i=1,j=1,k=0}^{Z} x_{ijk}}, \tag{1}$$

where Z is the number of zones. The distribution of these frequencies is visualized in Fig. 3d. To assess the trip chaining framework performance, we compared the observed (p_{ijk}) and estimated ($\hat{p}_{ijk}$) frequencies at the zone level.

Passenger Typology Analysis

The goal of the typology analysis was to understand the spatiotemporal trip patterns of passengers in the network and to use this to inform the trip chaining framework. First, we developed spatiotemporal passenger trajectories using mobile ticketing data to categorize passengers based on their network usage patterns. We constructed these trajectories, spanning July 2020–March 2021, in 10-min intervals, capturing each passenger's complete travel patterns. The entire set of trajectories is represented as the tensor $\mathcal{Y} \in \mathbb{R}^{N \times T \times L}$, where N represents the number of passengers, T represents the number of 10-min time intervals within the study period, and L represents the number of features (longitude and latitude). The trajectory tensor $\mathcal{Y}$ can also be considered as a panel of two matrices $\boldsymbol{Y}_1, \boldsymbol{Y}_2 \in \mathbb{R}^{N \times T}$, where rows correspond to unique user IDs and columns represent time intervals. Matrix elements contain either the longitude or latitude of the first ticket activation within a given 10-min period, or a zero in case of no activation. Figure 4 shows the trajectories of three randomly selected users throughout the study duration. These visualizations highlight the variable nature of individual mobility, as indicated by the diverse trajectory lengths and spatial distribution.

We then used dynamic time warping (DTW) (Vintsyuk 1972; Mueen et al. 2016) to compute pairwise distances between passenger trajectories in both the longitude and latitude dimensions. The DTW distance D_{ij} between the trajectories of passengers n and n' is given by

$$D_{ij} = \sum_{l=1}^{L} \frac{1}{C} \sum_{k=1}^{M} c_k \min_{\phi_k} d(\phi_k(\boldsymbol{y}_{n,l}, \boldsymbol{y}_{n',l})), \tag{2}$$

where C is a normalization constant, c_k is the weighting coefficient at each step, and $d(\cdot)$ is the distance metric, which in our case was Euclidean. $\boldsymbol{y}_{n,l}$ and $\boldsymbol{y}_{n',l}$ are the respective time series components (longitude or latitude) of the trajectory matrices $\boldsymbol{Y}_n$ and $\boldsymbol{Y}_{n'}$ of passengers n and n'. The warping path ϕ_k indicates the mapping of observations in $\boldsymbol{y}_{n,l}$ to those in $\boldsymbol{y}_{n',l}$.

Using the matrix of pairwise distances among all the users N, we applied the Ward method (Ward 1963) to cluster the passengers. The optimization objective for merging clusters $\mathcal{C}$ and $\mathcal{C}'$ is given by

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

where $|\mathcal{C}|$ and $|\mathcal{C}'|$ are the sizes of the clusters being merged, and $\boldsymbol{c}_{\mathcal{C}}$ and $\boldsymbol{c}_{\mathcal{C}'}$ are the respective cluster centroids. We refer to the clusters obtained as passenger *types* t, which comprise

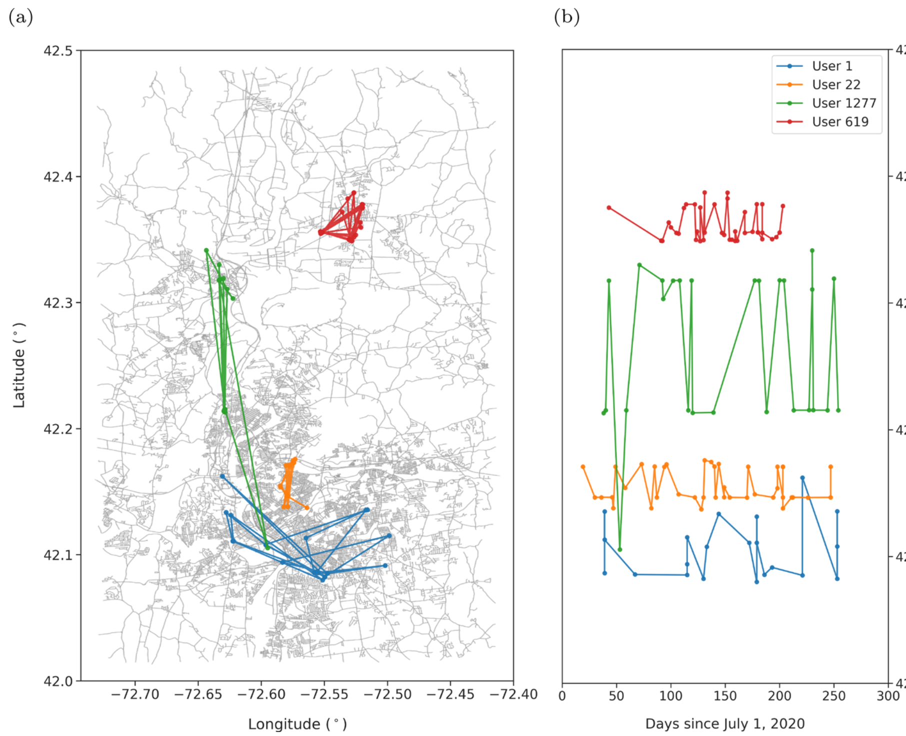

Fig. 4 Spatiotemporal trajectories of four randomly selected users from July 2020 to March 2021: **a** map showing the spatial movement of users between activation points, and **b** same user trajectories with latitude on the *y*-axis and days passed since July 1, 2020, on the *x*-axis to show the temporal variation in travel patterns

the spatiotemporal passenger *typology*. The procedure is summarized in Fig. 5.

Beyond providing insights into the overarching passenger travel patterns in the network, the resulting passenger types (typology) provide specific seasonal (c) type-specific (t) trip chaining parameters for the boarding (B), alighting (A), and transfer (T) distance thresholds, denoted as $\rho_{B,c,t}$, $\rho_{A,c,t}$, and $\rho_{T,c,t}$, respectively. These parameters are directly used in the trip chaining algorithms to determine the likelihood of boarding, alighting, and transferring at specific locations. For each identified passenger type, the parameters adjust the model to better reflect the observed travel behaviors. For instance, a commuter type may have shorter boarding distance thresholds due to more predictable and frequent travel patterns, while an irregular passenger type may have larger thresholds due to varied travel behaviors. Integrating these parameters into the framework can, therefore, accurately reflect the diverse travel behaviors observed throughout the year.

Type Classification

Using spatiotemporal and faretype features, we estimated models to then classify the passengers into the types obtained from the typology analysis. The aim was to validate the unsupervised typology grouping and ultimately develop an approach for inferring the passenger type of future users without having to repeat the computationally expensive clustering procedure. We also used this approach to determine passenger types in the data collected between March 2021 and December 2021.

The first step involved feature engineering to create a dataset with a holistic view of each user's travel pattern. This process had three main aspects: First, we captured patterns in ticket activations, including frequency, peak usage times, and statistics, such as the average and standard deviation of time intervals between activations. Second, we divided bus stop locations into 20 zones using K-means clustering and categorized each user's activations into these zones. Finally, we generated the usage proportion of different programs and fare types per user. The resulting dataset comprised 140 features and included data from all users up to March 2021, totaling 3,184 users. We used this dataset to train a classification model to classify all remaining users into passenger types. We used a training and test split of 60% and 40%, respectively, to evaluate the model's performance.

To identify the most effective classification model, we tested several well-known algorithms. These included

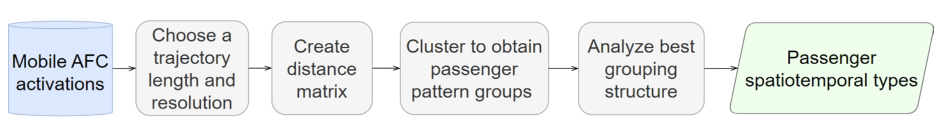

Fig. 5 Flowchart of the passenger typology analysis procedure

random forests (Breiman 2001), gradient boosting (GBM) (Friedman 2001), logistic regression (Berkson 1944), multilayer perceptron (MLP) (Minsky and Papert 2017), K-nearest neighbors (KNN) (Fix and Hodges 1989), and naive Bayes (Bayes and Price 1997). We evaluated each model based on its ability to accurately classify users into predefined types. We also evaluated the model based on other performance metrics to understand the strengths and weaknesses of each approach.

Trip Chaining

We designed a trip chaining framework that leverages mobile ticketing data. Although this data capture GPS coordinates when a passenger boards, it does not specify the bus stop ID. Additionally, it does not capture any information when passengers alight. This framework (Fig. 6) identifies boarding, alighting, and transfer stops to infer complete passenger trips, thereby allowing agencies to understand mobility patterns for better transportation planning. It is represented as

$$\hat{\mathbf{R}} = \mathbb{TC}(\mathbf{A};\boldsymbol{\theta}_m), \tag{4}$$

where $\mathbb{TC}$ is the trip chaining algorithm, which involves boarding ($\mathbb{BI}$), alighting ($\mathbb{AI}$), and transfer inferences ($\mathbb{TI}$), $\mathbf{A}$ is a matrix of activations with features such as coordinates, timestamps, and user IDs, and $\hat{\mathbf{R}}$ is an extended matrix with inferred trips including boarding stops, alighting stops, and a binary transfer indicator. The parameter vector $\boldsymbol{\theta}_m = [\kappa_m, \tau_m]$ comprises the distance threshold quantile κ_m and the transfer time threshold τ_m for a given season length m. We identify the best parameters by adjusting m to determine the ideal season duration m^*. Then, we optimize $\boldsymbol{\theta}_m$ to $\boldsymbol{\theta}_{m^*} = [\kappa^*_{m^*}, \tau^*_{m^*}]$ across all seasons. Here, $\kappa^*_{m^*}$ and $\tau^*_{m^*}$ denote the optimal values of κ and τ respectively, for the optimal season length m^*.

The trip chaining framework sequentially estimates the necessary stops and transfers from user activity data. Initially, boarding inference identifies probable boarding points and bus trips for each user. Following this, alighting inference estimates the alighting stops based on the time and location of the next boarding. Finally, transfer inference estimates possible transfers by analyzing the time and distance between an alighting and the subsequent boarding. These algorithms are described in the following sections. For the pseudocode of each algorithm, see "Appendix 2".

Boarding Inference

The boarding inference algorithm uses the activation locations from the mobile device's GPS to infer passenger boarding stops. We assigned activations to the nearest bus stop if the activation location and time were within an acceptable distance ($\rho_{B,c,t}$) and aligned with General Transit Feed Specification (GTFS) schedule data, considering a 15-min buffer to accommodate bus schedule deviations (Mohammed and Oke 2023). We also captured and linked additional details to each activation with a successful assignment, such as all possible bus trip IDs, to be used in the alighting inference step. (The detailed procedure is outlined in Algorithm 1, "Appendix 2").

Alighting Inference

Using the inferred possible boarding locations and GTFS data, we developed an alighting inference algorithm to predict alighting locations. First, we identified feasible stops between each pair of consecutive boardings using a distance threshold $\rho_{A,c,t}$ for walking from an alighting stop to the following boarding stop. Then, we estimated the alighting points based on proximity to the subsequent boarding location in the passenger's trajectory, as well as the feasibility of the trip happening based on GTFS data. We treated special cases, such as the last trip of the day, by assuming the day's final destination was the first boarding stop of the same day. Additionally, we excluded days with single activations due to inferential limitations. (Please refer to Algorithm 2, "Appendix 2" for further details).

Transfer Inference

Transfer inference assesses passenger trip legs, assigning a binary label denoting a "transfer" or "no transfer" based on time intervals and distances between consecutive trip legs. This process includes the consideration of a transfer time

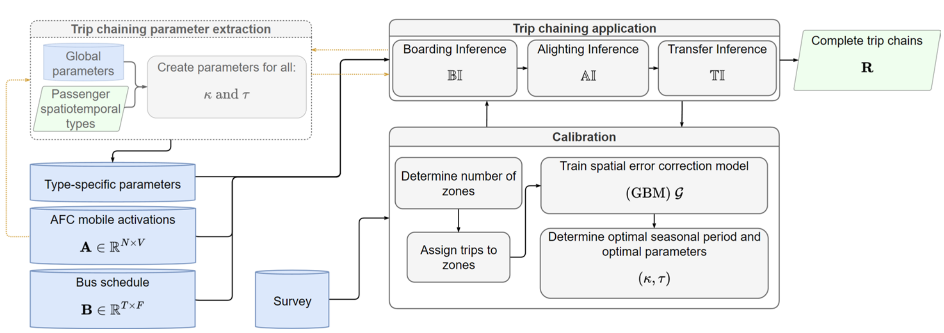

Fig. 6 Flowchart of the trip chaining framework

threshold, τ, which determines the maximum allowable time between consecutive trips to be considered as a continuous journey. This threshold is used in distinguishing between a transfer and two separate trips. (Details can be found in Algorithm 3, "Appendix 2").

Calibration

In the trip chaining process, several potential sources of error may arise. Travel surveys provide general location information but lack specific stop IDs, necessitating approximations through the use of Fuzzy matching and Levenshtein distance to align survey responses with specific bus stops. This potentially introduces inaccuracies in the data used to calibrate the model. Additionally, the accuracy of GPS data collected from mobile ticketing can vary due to signal interference, urban canyons, and other factors, affecting the precision of inferred boarding and alighting locations. Furthermore, the 10-min intervals used to construct passenger trajectories may not capture short-duration trips accurately, potentially leading to misclassification of boarding and alighting events. Finally, the inference of transfer points relies on the time and distance between consecutive boardings, and any inaccuracies in these measurements can lead to incorrect identification of transfer points.

To mitigate these potential errors and ensure the model's robustness, we employed two steps. First, we calibrated the model using the survey data, fine-tuning its parameters to best align with the observed boarding and alighting patterns. Given that the survey data only provide approximate locations, we aggregated stops into zones to account for potential inaccuracies in stop identification. Second, we validated the model's performance using a held-out portion of the mobile ticketing dataset not used in the calibration process. This validation step assessed the model's ability to accurately predict trips based on mobile ticketing data alone.

Our calibration approach first involved the aggregation of observations from the stop level to the zone level. Following this, we used a grid search to find the optimal parameters to fit trip chaining predictions of the zone-based trip frequencies to those observed from the survey conducted in the case study area. The key parameters included the seasonal period m and the type- and season-specific quantile κ_m and transfer time τ_m thresholds. We also incorporated a model to learn the spatial error structure of initial predictions and then correct them with predicted residuals. These steps are detailed in the following subsections.

Zonal Aggregation

Considering the limitations in sample size, computational resources, and validation requirements, we opted for zonal rather than stop-level analysis. This involved using K-means clustering to group stops based on their geographical coordinates. To determine the most representative spatial distribution of stops with efficient computational performance, we examined various zone counts, $Z \in \{5, 10, \dots, 70\}$. Our primary metric was the within-cluster sum of squares (WCSS), where a lower WCSS signifies data points more closely clustered around their respective centroids, indicating better clustering.

We identified the optimal number of clusters using the elbow heuristic, which balances model fitness (as indicated by WCSS) and complexity (reflected in the number of clusters). This method led to the selection of $Z = 20$ as the most suitable number of zones for aggregating trip origins and destinations in our trip chaining framework (see Fig. 7a). Figure 7b illustrates the spatial distribution of these zones and their encompassed stops.

Seasonal Period Selection

We proceeded to identify the optimal seasonal period length by partitioning the activation matrix $\mathbf{A}$ into several seasonal periods $m \in \{0, 1, 2, 3, 4\}$ months, with $m = 0$ representing the absence of seasonality. We then assessed each seasonal period's impact on key error metrics to identify the optimal duration (m^*) that yielded the lowest error values, indicating the most accurate modeling of seasonal variations.

This seasonality is incorporated into the framework by adjusting the distance thresholds for boardings, alightings, and transfers for each season c based on the identified optimal seasonal period (m^*). This means that for each season, the model uses the distance thresholds that reflect that specific season's travel patterns. During the boarding inference, for instance, the seasonal type-specific boarding distance threshold ($\rho_{B,c,t}$) accommodates seasonal variations in travel behavior, such as reduced travel distances during winter months. This process ensures that the trip chaining model dynamically adapts to seasonal changes, thereby improving prediction accuracy.

Type- and Season-Specific Parameter Selection

After establishing the optimal number of zones and seasonal period, we focused on optimizing the other parameters. First, for each season c and type t, we used global parameters to apply the trip chaining process. Then, the distance distributions for boarding, alighting, and transfers were determined using these preliminary inferred locations.

We then calibrated the quantile hyperparameter κ_{m^*}, which is used to calculate the distance thresholds for boarding $\rho_{B,c,t}$, alighting $\rho_{A,c,t}$, and transfers $\rho_{T,c,t}$, to optimize the trip chaining accuracy. This calibration was achieved by performing the trip chaining with different quantiles $\kappa_{m^*} \in \{0.40, 0.55, \dots, 0.95\}$ and then evaluating the results

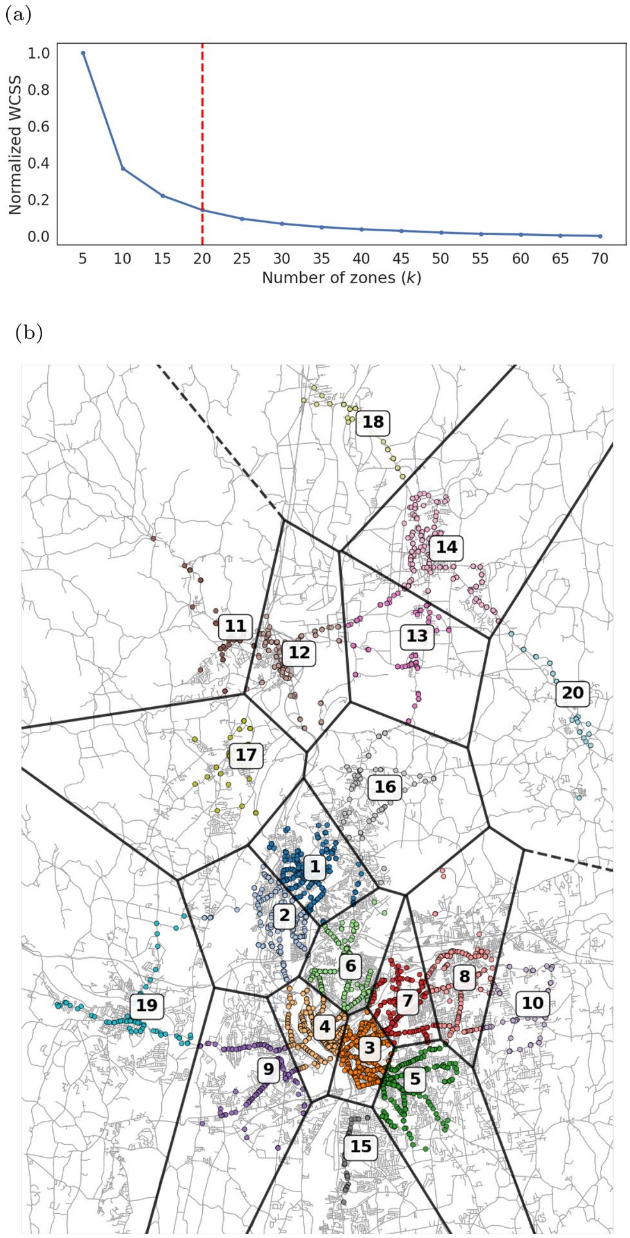

Fig. 7 Analysis of zones in the Pioneer Valley bus network: **a** the impact of varying the number of zones (Z) on the WCSS metric, which is normalized to facilitate comparison across the different number of zones (the selected value of $Z = 20$ represents an optimal point), and **b** spatial distribution of zones and stops

against the PVTA survey data using specified error metrics. We calculated the thresholds as

$$\rho_{\varphi,c,t} = F^{-1}_{\varphi,c,t}(\kappa), \tag{5}$$

where φ indicates boarding (B), alighting (A), or transfer (T), c is the season, t is the type, $F^{-1}_{\varphi,c,t}$ is the inverse cumulative distribution function of the distances, and κ is the quantile.

The calibration of τ_{m^*} follows a similar approach using $\tau_{m^*} \in \{5, 10, \dots, 60\}$, with a focus on optimizing the transfer inference accuracy. After finding the optimal $\kappa^*_{m^*}$ and $\tau^*_{m^*}$, they remain constant, but the distance parameters $\rho_{\varphi,c,t}$ vary for each season c and type t.

Spatial Error Learning

To enhance the accuracy of our trip chaining framework, we used a gradient boosting machine (GBM) model to learn and correct spatial errors in the trip chaining predictions. We trained this model on a full year (from July 2020 to July 2021) of AFC data to learn error structures across zones, using the residual r_{ijk}, which is the difference between observed survey frequencies p_{ijk} and predicted frequencies $\hat{p}_{ijk}$

$$r_{ijk} = p_{ijk} - \hat{p}_{ijk}. \tag{6}$$

We performed a grid search to obtain the best hyperparameters for the GBM model, denoted as $\mathcal{G}$. To mitigate the risk of overfitting in in the GBM, we employed fivefold cross-validation to compute the errors for each point searched. Figure 19 in "Appendix 5" shows the results of the grid search.

The GBM model's predicted residuals $\hat{r}_{ijk}$ were added to the initial trip chaining predictions $\hat{p}_{ijk}$ to produce the final adjusted trip frequencies $\tilde{p}_{ijk}$. This ensured that the model corrects for any spatial inaccuracies identified during the training phase

$$\hat{r}_{ijk} = \mathcal{G}(i, j, k) \tag{7}$$

$$\tilde{p}_{ijk} = \hat{p}_{ijk} + \hat{r}_{ijk}. \tag{8}$$

We validated this error-learning approach using 5 months of test data (from August 2021 to December 2021).

Performance Metrics

To assess the performance of the trip chaining framework, we quantified the closeness between the predicted zone-level trip frequencies and those derived from the survey. We used three key metrics for this purpose: mean absolute error (MAE), Jensen-Shannon divergence (JSD), and the symmetric mean absolute percentage error (SMAPE). These metrics are defined in "Appendix 3". Collectively, they offer a robust approach to assess the performance of the trip chaining model, ensuring that it accurately reflects the travel demand patterns observed in the bus transit network.

Results and Discussion

Spatiotemporal Passenger Types

We derived the optimal number of clusters by examining the dendrogram resulting from the Ward method for distinct and interpretable groupings, and by analyzing the silhouette metric, CH index, and WCSS across various cluster numbers. (These analyses are detailed in "Appendix 4".) Thus, we obtained four spatiotemporal passenger types, each with unique travel behaviors and patterns: Commuters (T1), Flex-Morning (T2), Afternoon-Shift (T3), and Irregulars (T4). A summary of type characteristics is given in Table 1.

Distinct travel rhythms among different user types are discovered from the patterns of hourly activations and the distribution of average intervals between these activations. The boxplots in Fig. 8a highlight the hourly concentration of travel activations for each type, from which we can determine the peak periods of use and the regularity of travel patterns. Similarly, Fig. 9b illustrates the distribution of average time intervals between activations for each user, highlighting the frequency and regularity of travel patterns in each

Table 1 Summary of distinct characteristics of identified spatiotemporal passenger types

Characteristic	Commuters (T1)	Flex-morning (T2)	Afternoon-shift (T3)	Irregulars (T4)
Number of unique passengers	218	836	706	1441
Average monthly boardings	1280	1750	260	490
Average daily boardings per user	5.8	2	0.36	0.34
Mean interval between activations (h)	16	33	30	48
Median time interval (h)	15	30	18	60
Modal interval between activations (h)	[9–10]	[12–13]	[3–4]	[17–18]
Travel frequency	High	Variable	Consistent	Sporadic
Probable travel purpose	Work/school	Work (variable shifts)	Work (fixed PM shift)	Various
Time of travel	Morning/Evening peak	Morning peak	Afternoon	Varied
Regular routes	Yes	Mostly	Mostly	No
Probable passenger demographics	Diverse	Workers, E/D	Workers	Students/shoppers
Spatial distribution	Urban centers	Urban/ suburban	Urban	Diverse

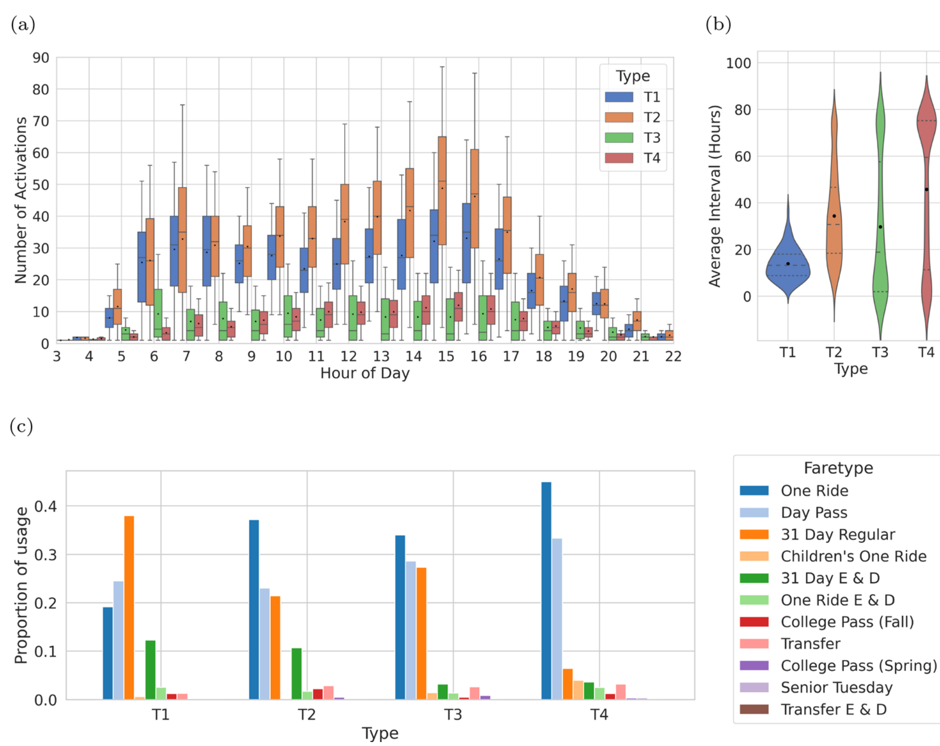

Fig. 8 Passenger type patterns: **a** boxplots of hourly user activations; dots indicate means, while horizontal lines indicate medians. **b** violin plots of average passenger time interval between activations; horizontal lines indicate first, second and third quartiles, while dots represent means, and **c** faretype usage distribution for each passenger type

type. These elements provide a detailed view of each group's travel interval symmetry and spread. For instance, T1 and T2 demonstrate pronounced morning–afternoon travel peaks (based on hourly activation counts), suggesting regular commutes for work or school. However, T1, with a lower number of unique passengers (218) and higher average monthly boardings (1280), appears to represents traditional commuters with consistent daily schedules and shorter mean intervals between activations at 16 h. In contrast, T2, which has a larger passenger base (836), less frequent daily boarding per user (2 trips), and longer mean interval between activations at 33 h, indicated a more diverse group of passengers that may include workers with variable shifts or other reasons for irregular commute times.

Simultaneously, T3 and T4 displayed unique travel patterns that diverged from the morning-afternoon travel peaks. T3 showed mostly consistent activations with low average monthly boardings (260), implying a less regular afternoon shift or other similar commitments. Alternatively, T4's plot of average time intervals between activations showed a wide distribution and longer mean interval (48 h), indicating that this group consisted predominantly of passengers with sporadic travel patterns and varied purposes for their journeys.

The spatial distribution of the passenger types, as shown in Table 1, also offers significant insights. T1 predominantly uses urban center stops, indicating their reliance on public transit for commuting within densely populated areas. T2 passengers are mostly seen in both urban and suburban areas, suggesting a blend of travel needs across different regions. T3 users also primarily utilize urban stops, similar to T1, possibly reflecting a similar work-related travel pattern but at different times. T4 with their diverse travel purposes shows a varied spatial distribution, utilizing stops across urban, suburban, and other areas, reflecting the diverse nature of their trips.

Faretype usage patterns (Fig. 8c) provided deeper insights into each passenger type's characteristics. For example, faretype analysis revealed T1's reliance on regular fares and T4's preference for one-ride and day passes, aligning with their respective inferred commuter and non-commuter behaviors. This faretype analysis, combined with the spatial and temporal patterns, provides a comprehensive understanding of the different passenger types within the study network. (Further details on the typology characterizations are also available in Abdalazeem and Oke 2023.)

Classification Model Assessment and Selection

We trained six models to predict passenger types obtained from the unsupervised typology. We evaluated the performance of each tested model by the following metrics: accuracy, precision, recall, F1 score, and the area under the Receiver-Operating Characteristics curve (AUC-ROC). These are summarized in Table 2. Accuracy measures the overall correctness of a model. Precision indicates the proportion of positive identifications that were correct. Recall, also known as sensitivity, measures the proportion of actual positives correctly identified. F1 Score provides a balance between precision and recall, important when uneven class distribution exists. AUC-ROC assesses a model's ability to distinguish between classes.

The gradient boosting type classifier showed superior performance, especially in accuracy (86.65%) and AUC-ROC (96.74%). The feature importance and confusion matrix plots for the model are shown in Fig. 9a, b. This result indicates that the spatiotemporal passenger types can be reasonably predicted from passenger features and thus validates the typology approach. This model can also be used to classify new users as more data become available and speed up the process of updating trip chaining results. This ensures that the bus network can dynamically adapt to changing passenger needs and preferences.

Optimal Seasonal Period and Type-Specific Parameters

We calculated the metrics using a grid search across various combinations of the seasonal period m, then across combinations of the transfer time threshold τ_{m^*} and the quantile κ_{m^*}, to identify the ones that minimized the MAE and JSD metrics, thus enhancing the model's precision. Figure 10 focuses only on seasonal periods to simplify presentation. It displays the minimum error value achievable within each seasonal period, derived from all possible combinations of

Table 2 Performance comparison of models for passenger type classification

Model	Accuracy	Precision	Recall	$F1$ score	AUC-ROC
Gradient boosting	**0.8666**	**0.8713**	**0.8666**	**0.8650**	**0.9674**
K-nearest neighbors	0.5714	0.5625	0.5714	0.5612	0.7972
Logistic regression	0.8140	0.8128	0.8140	0.8105	0.9433
Multi-layer perceptron	0.6523	0.5238	0.6523	0.5656	0.8399
Naive bayes	0.3980	0.5413	0.3980	0.3748	0.7987
Random forest	0.8422	0.8473	0.8422	0.8391	0.9528

The row in bold highlights the selected model

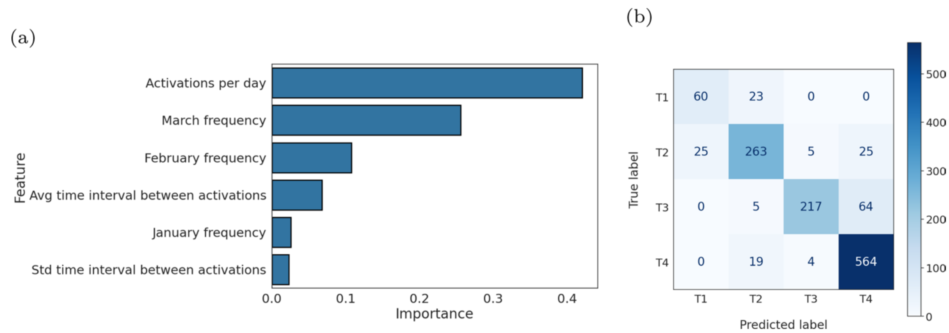

Fig. 9 Visualization of the gradient boosting type classification model performance: **a** feature importance, and **b** confusion matrix

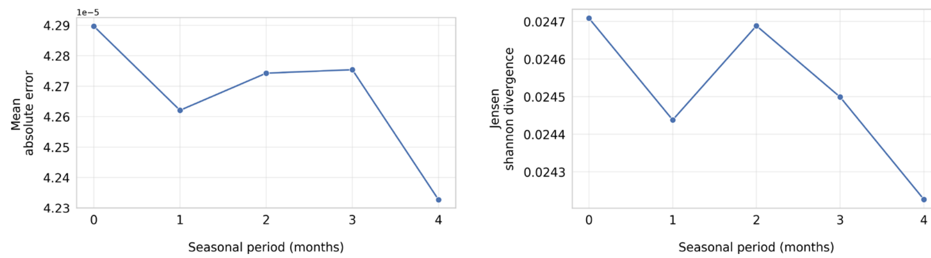

Fig. 10 Error metrics across different seasonal periods. The values reported represent the minimum error values derived from all combinations of seasonal periods and trip chaining parameters

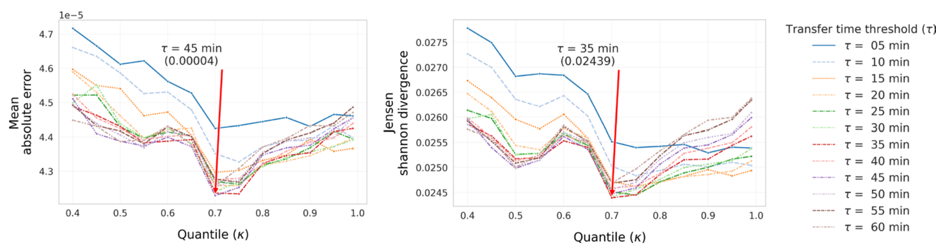

Fig. 11 Error trajectories for different combinations of transfer time threshold (τ_{m^*}) and quantile (κ_{m^*}). The red lines denote the optimal parameter settings that minimize the respective error metrics, with annotations indicating the optimal $\tau^*_{m^*}$ and the corresponding error value

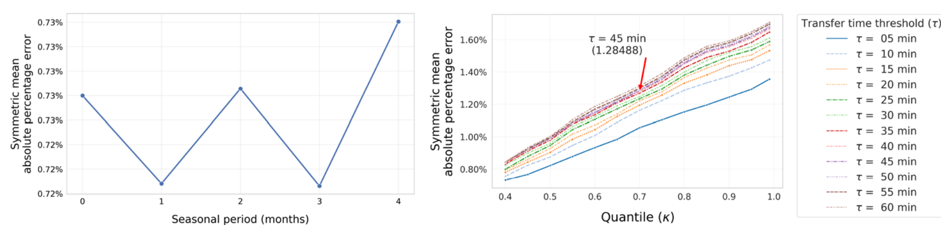

Fig. 12 SMAPE metrics across different seasonal periods and parameter combinations. The red line (right figure) denotes the SMAPE value obtained by the MAE and JSD-optimal values ($\tau_{m^*} = 45$ and $\kappa_{m^*} = 0.7$)

the seasonal period and the trip chaining parameters (κ_m, τ_m). We observed that extending the seasonal period to $m^* = 4$ months led to the lowest error values.

We conducted further examination of the specific impact of the trip chaining type-specific seasonal parameters within the optimal period of 4 months, as shown in Fig. 11. A clear trend was shown in the figure: as κ_{m^*} increased, the error rates initially decreased, indicating an improvement in model performance. Then, as κ_{m^*} increased further, the error rates worsened. We observed that the lines representing the 35-min and 45-min transfer time thresholds, τ_{m^*}, were lowest at $\kappa_{m^*} = 0.7$. Although shorter transfer times might be intuitively preferred for faster travel, the model does not substantially favor the 35-min threshold over the 45-min one, in terms of accuracy. Therefore, we chose the 45-min threshold, which provides a more conservative estimate of transfer locations.

To further validate our findings, we also examined the symmetric mean absolute percentage error (SMAPE) for various seasonal periods and parameters (Fig. 12). The differences in SMAPE values across the seasonal periods investigated were all less than 1%, which validates our selection of $m^* = 4$. Additionally, the SMAPE of the MAE and JSD-optimal values ($\tau_{m^*} = 45$ and $\kappa_{m^*} = 0.7$) had a slightly higher SMAPE value of 1.28%, which also indicates good performance. The small differences in SMAPE values indicate that while there is a slight trade-off, the overall model performance remains robust and reliable with the chosen parameters. Thus, we chose to prioritize the parameters optimized for MAE and JSD due to their consistency in minimizing other error metrics and their overall better alignment with the observed travel patterns.

Using these optimized parameters, we calculate the three distance thresholds (boarding $\rho_{B,c,t}$, alighting $\rho_{A,c,t}$, and transfer $\rho_{T,c,t}$) for each passenger type t and season c and show the cumulative distribution function (CDF) plots in Fig. 13. In the summary table (Table 3), we present both the actual distance thresholds and their respective percentage changes from the baseline season (Jan–April) for each passenger type and seasonal period. The boarding distance thresholds, $\rho_{B,c,t}$, show minor fluctuations across different types and seasons. For example, the percentage change in T1 from the baseline season shows a decrease of 7.41% in May–August (S2) to a decrease of 5.76% in Sept–Dec (S3). These variations, though modest, could indicate slight shifts in boarding behaviors or operational adjustments during different seasons.

The alighting distance thresholds, $\rho_{A,c,t}$, display more pronounced changes, especially in T4 and T3. In T4, there is a significant increase in later seasons, with a rise of 29% in May–August and 15% in Sept–Dec from the baseline season. Conversely, T3 experiences a marked decrease with

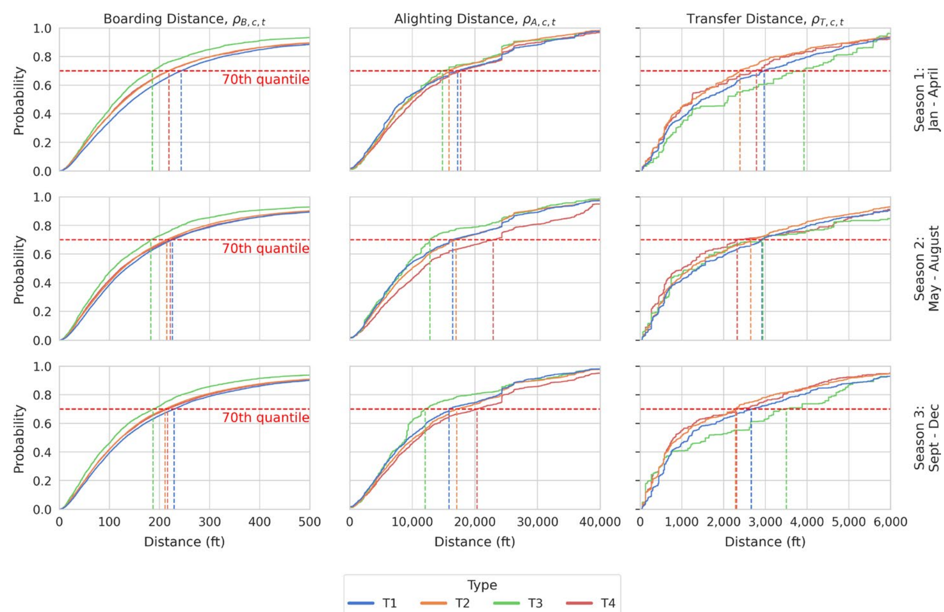

Fig. 13 Cumulative distribution function (CDF) plots showing the distance thresholds for different passenger types t for each season c. For boarding, these CDFs show distances between activation locations and their nearest bus stops. For alighting and transfers, the CDFs display the estimated distances between predicted alighting points and subsequent boarding stops, differentiating whether the stop was used for transferring or solely for alighting. Vertical lines correspond to the selected threshold based on the optimal $\kappa^*_{m^*}$ value of 0.7

Table 3 Summary table of distance thresholds and their percentage change from the baseline for each passenger type t and seasonal period c

Season	Type	Boarding distances $\rho_{B,c,t}$ (ft)		Alighting distance $\rho_{A,c,t}$ (ft)		Transfer distance $\rho_{T,c,t}$ (ft)	
		Value	% Change	Value	% Change	Value	% Change
Jan–April (S1)	T1	243	0.0	17,253	0.0	2086	0.0
	T2	218	0.0	15,871	0.0	2050	0.0
	T3	185	0.0	14,817	0.0	2035	0.0
	T4	218	0.0	17,736	0.0	2086	0.0
May–August (S2)	T1	225	−7.4	16,425	−4.8	1831	−12
	T2	214	−1.8	16,978	7.0	2086	1.8
	T3	182	−1.6	12,817	−14	1427	−30
	T4	221	1.4	22,901	29	1750	−16
Sept–Dec (S3)	T1	229	−5.8	15,871	−8.0	1736	−17
	T2	211	−3.2	17,105	7.8	1926	−6.0
	T3	187	1.1	12,050	−19	1452	−29
	T4	216	−0.9	20,342	15	2050	−1.7

a decline of 14% in May–August and 19% in Sept–Dec. These trends suggest a notable seasonal impact on passenger alighting behavior, potentially due to changes in passenger destinations or service routes during these periods.

Finally, we observed considerable variations in the transfer distance thresholds, $\rho_{T,c,t}$, particularly for T3, with a decrease of 30% in May–August and 29% in Sept–Dec from the baseline. Such significant changes indicate that the transfer process is substantially influenced by seasonal factors, possibly due to changes in passenger flow or operational adjustments during these periods.

Error Correction Model Performance

We performed a grid search to find the optimal GBM spatial error correction model hyperparameters that minimize the cross-validation root-mean-squared error (CV-RMSE). The search space included the number of estimators z, learning rate α, and maximum tree depth d. Detailed results and training trajectories of the grid search are provided in "Appendix 5".

Figure 14a presents a geographical heatmap, constructed based on the cumulative sum of errors for trips originating from each of the 20 zones. This heatmap shows the spatial distribution of the trip chaining model's residuals across the network, before applying the GBM spatial error correction, which allows for a visual assessment of the model's performance across the zones.

The predictive performance of the GBM spatial error corrector is visually represented in Figure 14b, which displays a log–log plot of predicted versus observed values. The model shows varying levels of accuracy across different ranges: for observed values lower than 10^{-3}, the model's predictions are less accurate, deviating from the ideal 45-degree line. However, as the observed values increase, the model's performance improves significantly, indicating that the GBM is more adept at predicting higher error values with greater precision.

The feature importances (Fig. 14c) indicate that the origin zone is the most relevant feature (0.62), followed

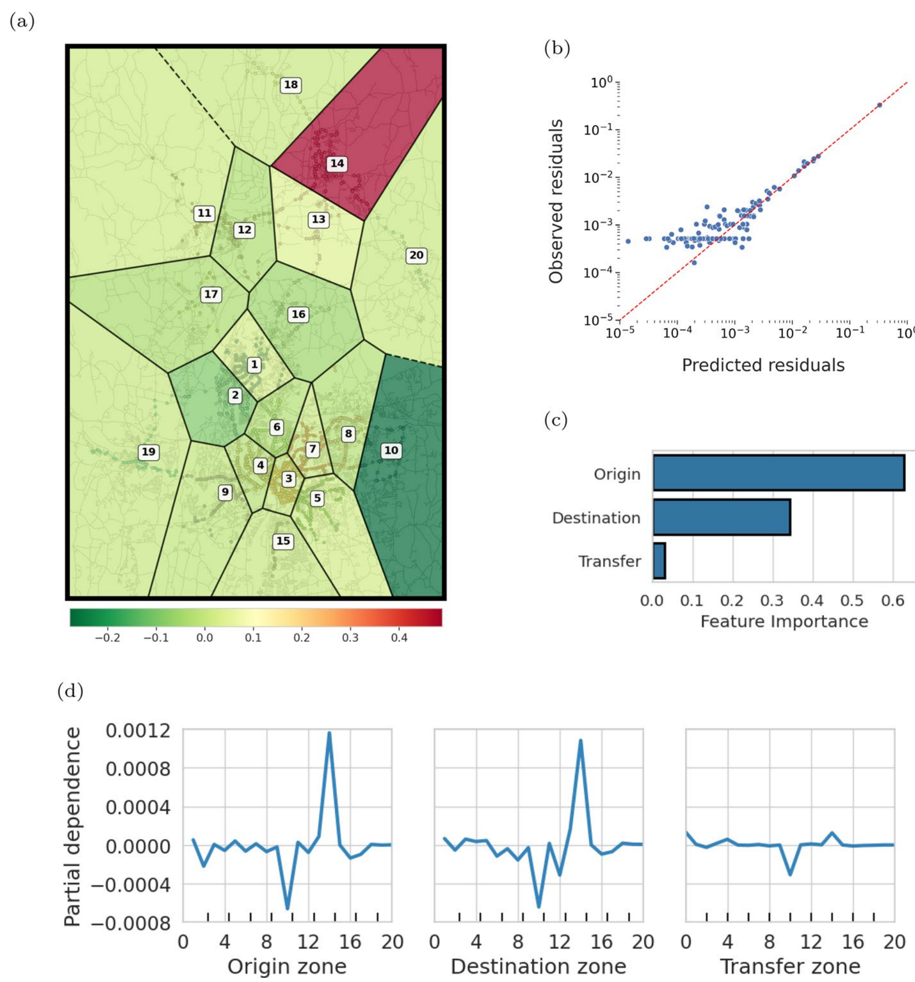

Fig. 14 Gradient boosting spatial error correction model: **a** geographical heatmap highlighting spatial residuals across 20 zones, **b** log–log plot of predicted versus observed residuals, showcasing model performance, **c** feature importance plot, and **d** partial dependence plots for origin, destination, and transfer zones

by the destination zone (0.34) and then the transfer zone (0.04). Also, for both origin and destination zones in the partial dependence plots (Fig. 14d), we observe that lower zone numbers (below 10), which represent the central urban core within the bus network, exhibit low values close to zero. This is supported by the heatmap in Fig. 14a, where these zones show minimal error, suggesting a high level of accuracy in model predictions for these areas in the original model. However, an exception is seen with zone 14, which includes a major university town to the north of the central urban core. Here, the plot shows a peak at 0.0012, and similarly, the heatmap indicates higher errors, suggesting an underestimation of travel flows by the model in this zone. This is to be expected, since transit in the whole zone is free and does not require ticket activations, unlike the other zones of the network.

In contrast, zone 10, which is a small area with few stops just outside the main central urban core, has a negative value of −0.0008 in the partial dependence plot and is also highlighted in the heatmap with a significant negative error. The overestimation of flows in zone 10 could be attributed to discrepancies in survey or AFC data. Since this zone is not as densely populated or transit-heavy as the central urban core, the AFC data might not be as robust or frequent, leading to potential overestimation in travel flow predictions.

Examining the transfer zones, the values remain mostly consistent and close to zero across all zones, which suggests that the framework is competent in predicting transfers. Nevertheless, a slight dip to −0.0002 is observed for zone 10, suggesting that the framework predicts slightly more transfers occurring in this zone than observed in the data.

This analysis is useful for identifying zones that are significant for predicting errors in travel flows. As seen from the plot, zones 10 and 14 appear to be influential for both origin and destination zones, while zone 10 is noteworthy for transfer zones. This information can help direct future data collection efforts, such as surveys, to improve the model performance.

Overall Trip Chaining Framework Performance

To assess the enhancements made to our trip chaining model, including the integration of seasonality, passenger typology, and spatial error correction, we compared its performance before and after these enhancements. Figure 15 illustrates this comparison. Figure 15a displays a wider spread of data points from the ideal line, indicating lower accuracy prior to the enhancements. Conversely, Fig. 15b demonstrates a significant improvement, with data points more closely aligned to the ideal line, indicating the model's improved accuracy.

Additionally, we validate our enhanced trip chaining model by comparing the trip frequencies derived from our framework using a test set of AFC observations (mobile activations) against those obtained from survey data. Figures 16a, b present heatmaps that show the correlation between the framework's predicted frequencies and the empirical survey data frequencies. The high similarity between the heatmaps in Figs. 16a, b, as well as the close alignment between observed and predicted trip frequencies in Fig. 15b confirm the high degree of accuracy achieved by our framework following the applied modifications.

The generalizability of our trip chaining model is shown by the comparison of performance metrics between the training and test sets, as shown in Table 4. The table shows that the model maintains a consistent level of accuracy when applied to unseen data, with only a slight increase in JSD and MAE from the train to the test set, by 24% and 26%,

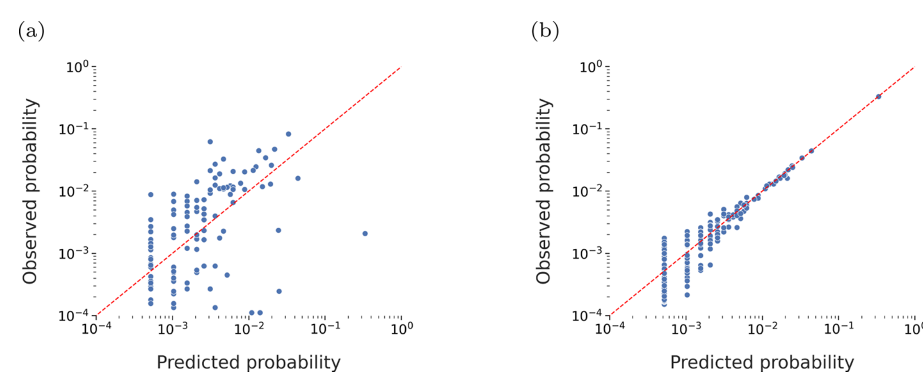

Fig. 15 Scatter plots of observed versus predicted trip frequencies in the trip chaining model before and after improvements: **a** model performance before enhancements showing a wider scatter from the line of best fit, and **b** model performance after applying seasonality, passenger typology, and spatial error correction showing tighter clustering around the line, indicating better model accuracy after the enhancements

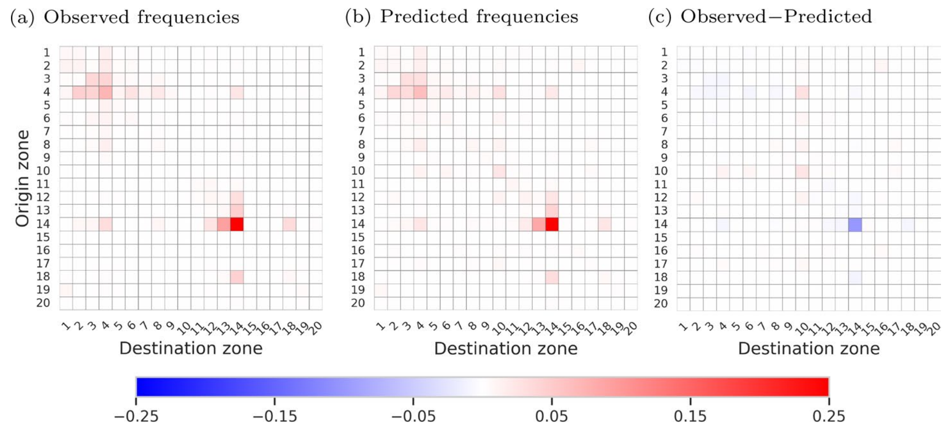

Fig. 16 Validation of model accuracy with the test set. Heatmaps of **a** observed trip frequencies (via survey), **b** predicted trip frequencies, and **c** errors between predicted and observed trip frequencies by zone

Table 4 Performance metrics of the model on the training and test sets

Dataset	MAE (10^{-4})	JSD	SMAPE (%)
Training	0.42	0.02	1.28
Test	**0.54**	**0.03**	**1.69**

The row in bold highlights the performance on the test set

respectively, as well as a modest increase of 32% (0.41 percentage points) in the SMAPE.

Finally, we evaluated the impact of incorporating seasonality, typology, and spatial error correction on our trip chaining framework's performance by testing various combinations of these elements. The resulting data, presented in Table 5, offer insight into how each innovation contributes to the overall effectiveness of the model.

The table shows that incorporating both seasonality and typology information (Model 5) led to a modest improvement in metrics: MAE improved to 1.51×10^{-4}, with a JSD of 0.28, and an SMAPE of 7.83%. This represents a minor enhancement over the base model with no enhancements (Model 1).

However, when we applied spatial error correction independently (Model 2), the results were significantly better, with reductions in MAE to 0.65×10^{-4}, JSD to 0.04, and SMAPE to 1.36%. This translates to approximately a 57% improvement in MAE, an 86% reduction in JSD, and an 84% improvement in SMAPE compared to the base model.

We saw the most impactful results when spatial correction was combined with seasonality and typology (Model 6). This integration yielded MAE value of 0.42×10^{-4}, a JSD of 0.02, and a SMAPE of 1.28%. The improvement here is substantial: about 34% better in MAE, 42% in JSD, and 6% in SMAPE compared to spatial correction alone. This highlights a key insight: while spatial error correction significantly enhances accuracy by itself, its effectiveness is further amplified when integrated with season-awareness and typology information.

Conclusion

In this study, we developed an innovative trip chaining framework tailored for transit systems that typically lack advanced data collection capabilities. This is particularly a significant issue in regional transit networks, which make up a considerable portion of U.S. transit infrastructure.

Table 5 Training performance metrics across different models with varying combinations of seasonality, typology analysis, and spatial error correction

Model	Season-aware	Typology-informed	Spatial error corrected	Performance metrics		
				MAE ($\times 10^{-4}$)	JSD	SMAPE (%)
1	–	–	–	1.52	0.29	8.43
2	–	–	✓	0.65	0.04	1.36
3	–	✓	–	1.54	0.28	7.92
4	–	✓	✓	0.54	0.03	1.29
5	✓	✓	–	1.51	0.28	7.83
6	✓	✓	✓	**0.42**	**0.02**	**1.28**

The row in bold corresponds to the performance of the model configuration that produced the best results

Our approach utilized mobile ticketing data, which is more accessible and cost-effective compared to the traditional smart card-based systems. We enhanced the precision of the trip chaining by extracting and analyzing travel patterns from this data. This approach involved a multi-step process: creating spatial trajectories for users, clustering these to identify distinct passenger behaviors, and then developing a framework for trip chaining. Central to this was the incorporation of typology and seasonality considerations, which allowed us to adapt the framework to varying travel patterns throughout the year. To refine our model's performance further, we integrated a gradient boosting machine, targeting those areas where data were unexplained or inconsistent.

We identified four distinct spatiotemporal passenger types: Commuters, Flex-Morning, Afternoon-Shift, and Irregulars, each exhibiting unique travel behaviors. Then, we estimated a gradient boosting machine (GBM) classifier with an accuracy of 86% which served to validate the typology. The classifier also facilitates the type prediction of future passengers to further streamline the trip chaining inference.

Integrating typology, seasonality, and spatial error correction into our trip chaining model significantly enhanced accuracy, especially through spatial error correction via gradient boosting. These improvements notably minimized prediction errors, evident in improved performance metrics: symmetric mean absolute percentage error (SMAPE), mean absolute error (MAE), and Jensen–Shannon divergence (JSD). We obtained the optimal results when all three enhancements—type-specific parameters, seasonality, and spatial correction—were applied together. This resulted in an overall enhancement of 72% in MAE, 93% in JSD, and 85% (7.15 percentage points) in the SMAPE, compared to the base model without these features.

The framework's ability to generate seed matrices for comprehensive origin–destination demand models can aid transit authorities in making data-driven decisions to improve service efficiency and respond to changing travel patterns. This is especially important in times of disruption, where accurate demand modeling can guide rapid adaptations to ensure continued mobility. Future research directions include integrating additional data sources, such as smart card data or GPS traces, to further enhance the model's accuracy and expand its potential applications. By bridging the data gap in regional transit systems, our framework offers a promising path toward a more efficient and responsive public transportation landscape.

Appendix 1: Symbols and Abbreviations

The acronyms and essential terms used throughout this paper are summarized in Tables 6 and 7, respectively.

Table 6 List of acronyms used in this paper

Acronym	Description
ADCS	Automated data collection systems
AFC	Automated fare collection
GBM	Gradient boosting machine
GPS	Global positioning system
GTFS	General transit feed specification
KNN	*K*-nearest neighbors
MLP	Multi-layer perceptron
OD	Origin–destination
ODX	Origin–destination-transfer
PVTA	Pioneer valley transit authority

Table 7 List of mathematical symbols used in this paper

Type	Symbol	Definition
Indices	F	Number of fields in the bus timetable (e.g., trip ID, route ID, stop ID, arrival/departure times)
	$i,j = 1,\dots,Z$	Origin and destination zones, respectively
	$k = 0,\dots,Z$	Transfer zones (0 indicates no transfer)
	L	Number of features (longitude and latitude)
	$n = 1,\dots,N$	Passengers
	T	Number of 10-min time intervals within the study period
	V	Number of features in the activation matrix (e.g., user ID, timestamp, coordinates)
	Z	Number of zones
	$\varphi = \{A, B, T\}$	Phase (alighting, boarding, transfer)
Sets	$c \in \mathcal{C}$	Season with optimal period m^*
	$m \in \mathcal{M}$	Seasonal period
	$s \in \mathcal{S}$	Network stop
	$t \in \mathcal{P}$	Passenger type within the spatiotemporal passenger typology
Scalars	D_{ij}	DTW distance between the trajectories of passengers n and n'
	p_{ijk}	Observed trip frequencies from origin i to destination j through transfer k
	r_{ijk}	Residual between the survey and the AFC-derived seed matrices
	x_{ijk}	Observed trip flows from origin i to destination j through transfer k
	δ_a	Time interval between alighting and next boarding
	δ_s	Shortest time interval between activation and incoming bus arrival
	$\rho_{\varphi,c,t}$	Distance threshold for phase φ, season c and passenger type t
Vectors	$\boldsymbol{c}_{\mathcal{C}}$ and $\boldsymbol{c}_{\mathcal{C}'}$	Centroids for clusters $\mathcal{C}$ and $\mathcal{C}'$, respectively
	$\lvert\mathcal{C}\rvert$ and $\lvert\mathcal{C}'\rvert$	Size of clusters $\mathcal{C}$ and $\mathcal{C}'$, respectively
	$\boldsymbol{y}_{n,l}, \boldsymbol{y}_{n',l}$	Time series components of the trajectory matrices for passengers n and n'
	$\boldsymbol{\theta_m} = [\kappa_m, \tau_m]$	Parameter vector of seasonal distance quantile κ_m and the transfer time threshold τ_m
	$\boldsymbol{\phi}_k$	Warping path indicating the mapping of observations in the time series $\boldsymbol{y}_{n,l}$ to those in $\boldsymbol{y}_{n',l}$
Matrices	$\mathbf{A} \in \mathbb{R}^{N\times V}$	Activation matrix
	$\mathbf{B} \in \mathbb{R}^{T\times F}$	Bus timetable
	$\hat{\mathbf{R}} \in \mathbb{R}^{N\times V+3}$	Matrix of completed trip chains/inferred trips for all users
	$\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{N\times T\times L}$	Spatiotemporal passenger trajectory tensor
	$\boldsymbol{Y}_1, \boldsymbol{Y}_2 \in \mathbb{R}^{N\times T}$	Matrices derived from the spatiotemporal passenger trajectory tensor $\boldsymbol{\mathcal{Y}}$, representing longitude and latitude data, respectively
Functions	$d(\cdot)$	Chosen distance metric (e.g., Euclidean)
	$\lvert \cdot \rvert$	Size of a cluster $\cdot$

Appendix 2: Trip Chaining Algorithms

The trip chaining algorithms for the boarding, alighting, and transfer inference are shown in Algorithms 1, 2, and 3, respectively.

Algorithm 1 Boarding Inference $\mathbb{BI}(\mathbf{A}, \mathcal{S}, \mathbf{B}; \boldsymbol{\theta}_{m^*})$

$\mathbf{A}' \leftarrow \mathbf{A}$ ▷ Create a copy of the activation data frame to add inferred boarding stops
$\mathcal{C} \leftarrow \{\text{Season } 1, \text{Season } 2, \text{Season } 3, \cdots\}$ within m^* ▷ Set of seasons within the optimal seasonal period m^*
for $c \in \mathcal{C}$ **do**
    $\boldsymbol{\theta}_{m^*} \leftarrow [\kappa^*_{m^*}, \tau^*_{m^*}]$
    $\rho_{B,c,t} = F^{-1}_{B,c,t}(\kappa^*_{m^*})$
    for $i = 1$ to $|\mathbf{A}|$ **do**
        $\mathcal{S}_i \leftarrow s \in \mathcal{S} : d(\mathbf{A}_i, s) \le \rho_{B,c,t}$ ▷ Find set of stops within allowed boarding radius for activation i
        if $\mathcal{S}_i = \varnothing$ **then**
            $\mathbf{A}'[b_i] \leftarrow \text{NULL}$ ▷ Assign no boarding stop for activation i
        else
            $\delta_s \leftarrow \min_{s \in \mathcal{S}_i}(t(\text{next bus arrival at } s) - \text{activation time})$ ▷ Find the smallest time interval in minutes between activation time and closest incoming bus at each stop in $\mathcal{S}_i$
            $s^* \leftarrow \text{argmin}_{s \in \mathcal{S}_i}(t(\text{next bus arrival at } s) - \text{activation time})$ ▷ Find the stop with the smallest time interval in minutes between activation time and closest incoming bus
            if $\delta_s > 15$ **then**
                $\mathbf{A}'[b_i] \leftarrow \text{NULL}$ ▷ Assign no boarding stop
            else
                $\mathbf{A}'[b_i] \leftarrow s^*$ ▷ Assign stop with δ_s as boarding stop
            end if
        end if
    end for
end for
return $\mathbf{A}'$

Algorithm 2 Alighting Inference $\mathbb{AI}(\mathbf{A}', \mathcal{S}, \mathbf{B}; \boldsymbol{\theta}_{m^*})$

$\mathbf{A}'' \leftarrow \mathbf{A}'$ ▷ Initialize a new data frame with inferred boarding stops
$\mathcal{C} \leftarrow \{\text{Season } 1, \text{Season } 2, \text{Season } 3, \cdots\}$ within m^* ▷ Set of seasons within the optimal seasonal period m^*
for $c \in \mathcal{C}$ **do**
  $\boldsymbol{\theta}_{m^*} \leftarrow [\kappa^*_{m^*}, \tau^*_{m^*}]$
  $\rho_{A,c,t} = F^{-1}_{A,c,t}(\kappa^*_{m^*})$
  for $i = 1$ to $|\mathbf{A}'|$ **do**
    $\mathcal{T}\nabla \leftarrow \{t \in \mathbf{B} \text{ at } \mathbf{A}'[b_i]\}$ ▷ Find possible bus trip IDs for boarding stop $\mathbf{A}'[b_i]$ from bus schedule $\mathbf{B}$
    $\mathcal{S}_i \leftarrow \{s \in \mathcal{S} : s \text{ visited by } t \text{ between activations } i \text{ and } i+1\}$ ▷ Find stops visited by trips $\mathcal{T}\nabla$ between current and next activation time
    $\mathcal{S}_i \leftarrow \{s \in \mathcal{S}_i : d \leq \rho_{A,c,t}\}$ ▷ Filter stops to retain ones within allowed alighting radius
    if $\mathcal{S}_i = \varnothing$ **then**
      $\mathbf{A}''[a_i] \leftarrow \text{NULL}$ ▷ Assign no alighting stop for activation i
    else
      $s^* \leftarrow \text{argmin}_{s \in \mathcal{S}_i} d(\mathbf{A}'_{i+1}, s)$ ▷ Find closest stop to next activation location
      $\mathbf{A}''[a_i] \leftarrow s^*$ ▷ Assign closest stop as alighting stop for activation i
    end if
  end for
end for
return $\mathbf{A}''$ ▷ Return the data frame with inferred alighting stops

Algorithm 3 Transfer Inference $\mathbb{TI}(\mathbf{A}''; \boldsymbol{\theta}_{m^*})$

$\mathbf{A}''' \leftarrow \mathbf{A}''$ ▷ Create a copy of the data frame with alighting stops
$\mathcal{C} \leftarrow \{\text{Season } 1, \text{Season } 2, \text{Season } 3, \cdots\}$ within m^* ▷ Set of seasons within the optimal seasonal period m^*
for $c \in \mathcal{C}$ **do**
  $\boldsymbol{\theta}_{m^*} \leftarrow [\kappa^*_{m^*}, \tau^*_{m^*}]$
  $\rho_{T,c,t} = F^{-1}_{T,c,t}(\kappa^*_{m^*})$
  for $i = 1$ to $|\mathbf{A}|$ **do**
    $d \leftarrow d(\mathbf{A}''[a_i], \mathbf{A}''[b_{i+1}])$ ▷ Find distance between current alighting stop and next boarding stop
    $\delta_a \leftarrow \mathbf{A}''[\text{time}_{i+1}] - \mathbf{A}''[\text{time}_i]$ ▷ Find time interval in minutes between alighting and next boarding
    if $d > \rho_{T,c,t}$ or $\delta_a > \tau^*_{m^*}$ **then**
      $\mathbf{A}'''[c_i] \leftarrow 0$ ▷ Assign as non-transfer trip for activation i
    else
      $\mathbf{A}'''[c_i] \leftarrow 1$ ▷ Assign as transfer trip for activation i
    end if
  end for
end for

$R \leftarrow \mathbf{A}'''$
return R ▷ Return the data frame with inferred transfer trips

Appendix 3: Error Metrics

Here, we provide a detailed explanation of the metrics used to evaluate the performance of the trip chaining model. These metrics include the mean absolute error (MAE), Jensen–Shannon divergence (JSD), and the symmetric mean absolute percentage error (SMAPE).

Appendix 3.1: Mean Absolute Error

The mean absolute error (MAE) provides a straightforward measure of the average magnitude of the errors, regardless of their direction. It is calculated as the mean of the absolute differences between the predicted and observed values

$$\mathrm{MAE} = \frac{1}{Z^2(Z+1)} \sum_{i,j,k} \left| p_{ijk} - \tilde{p}_{ijk} \right|. \tag{9}$$

Appendix 3.2: Jensen–Shannon Divergence

The Jensen-Shannon Divergence (JSD) (Lin 1991) provides a symmetric and bounded method to measure the dissimilarity between two probability distributions, P and $\tilde{P}$. It is defined as follows:

$$\mathrm{JSD}(P||\tilde{P}) = \frac{1}{2} D_{\mathrm{KL}}(P||M) + \frac{1}{2} D_{\mathrm{KL}}(\tilde{P}||M), \tag{10}$$

where M represents the mean distribution given by

$$M = \frac{1}{2}(P + \tilde{P}) \tag{11}$$

and $D_{\mathrm{KL}}(P||\tilde{P})$ is the Kullback–Leibler divergence given by

$$D_{\mathrm{KL}}(P||\tilde{P}) = \sum_{ijk} p_{ijk} \log \frac{p_{ijk}}{\tilde{p}_{ijk}}, \tag{12}$$

where p_{ijk} and $\tilde{p}_{ijk}$ denote the probabilities in the distributions P and $\tilde{P}$, respectively.

Appendix 3.3: Symmetric Mean Absolute Percentage Error

The symmetric mean absolute percentage error (SMAPE) is defined as a measure of prediction accuracy that normalizes the absolute error, providing a percentage-based metric (Jon Scott 1987). It is calculated as follows:

$$\mathrm{SMAPE} = \frac{100}{Z^2(Z+1)} \sum_{i,j,k} \frac{|p_{ijk} - \tilde{p}_{ijk}|}{|p_{ijk}| + |\tilde{p}_{ijk}|}. \tag{13}$$

SMAPE is closely related to the mean absolute percentage error (MAPE) but offers significant advantages for our data. While MAPE is straightforward, it can be heavily skewed by very small actual values, resulting in disproportionately large errors. In contrast, SMAPE is symmetric and bounded between 0 and 100%, which is useful for transit data where trip counts can vary significantly across zones. This makes SMAPE a more reliable and interpretable metric for evaluating the performance of our trip chaining framework.

Appendix 4: Spatiotemporal Passenger Types

The determination of the optimal number of clusters was conducted through an inspection of the dendrogram (Fig. 17), evaluation of cluster validity metrics, and the analysis of the within-cluster sum of squares (WCSS) via the elbow heuristic. We considered the distinctiveness and interpretability of the resulting groups, with as small a number of clusters as possible. Figure 18a–c display the silhouette metric, the Calinski–Harabasz (CH) index, and the WCSS for various numbers of clusters, respectively.

The silhouette score remained relatively consistent across different cluster solutions, with the first local maximum being at four clusters. The CH index showed a marked decrease when transitioning from four to five clusters. Similarly, the WCSS presents an elbow at four clusters (Fig. 18c), suggesting that optimal cluster homogeneity is achieved at this point. These observations, coupled with the imperative

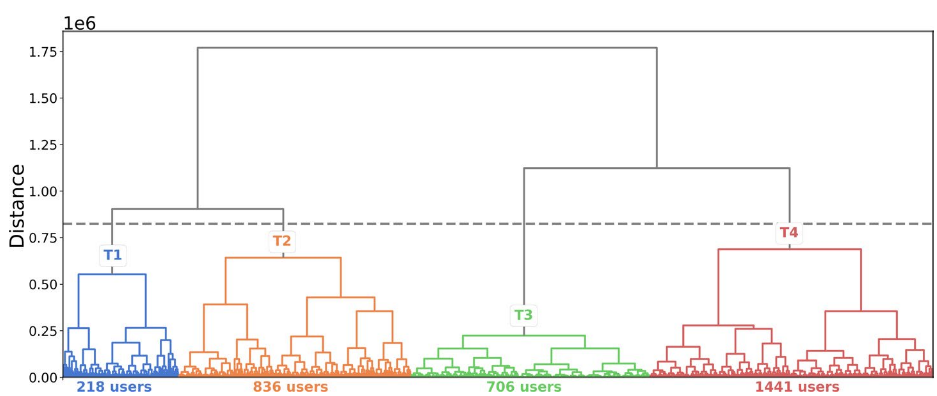

Fig. 17 Dendrogram representing the classification of spatiotemporal passenger types within the PVTA bus transit network. Each color corresponds to one of the four identified passenger types: T1, T2, T3, and T4

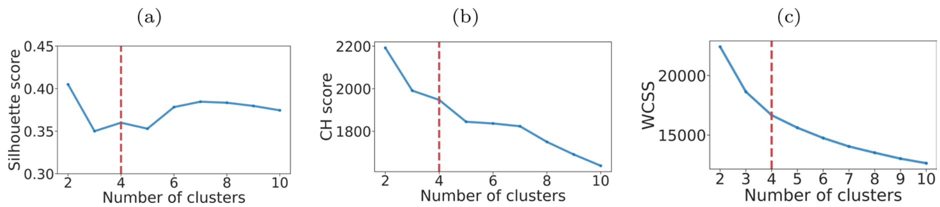

Fig. 18 Values of the **a** silhouette metric, **b** Calinski–Harabasz (CH) index, and **c** the within-cluster sum of squares for various cluster numbers. The red dashed line across all plots indicates the chosen optimal number of clusters

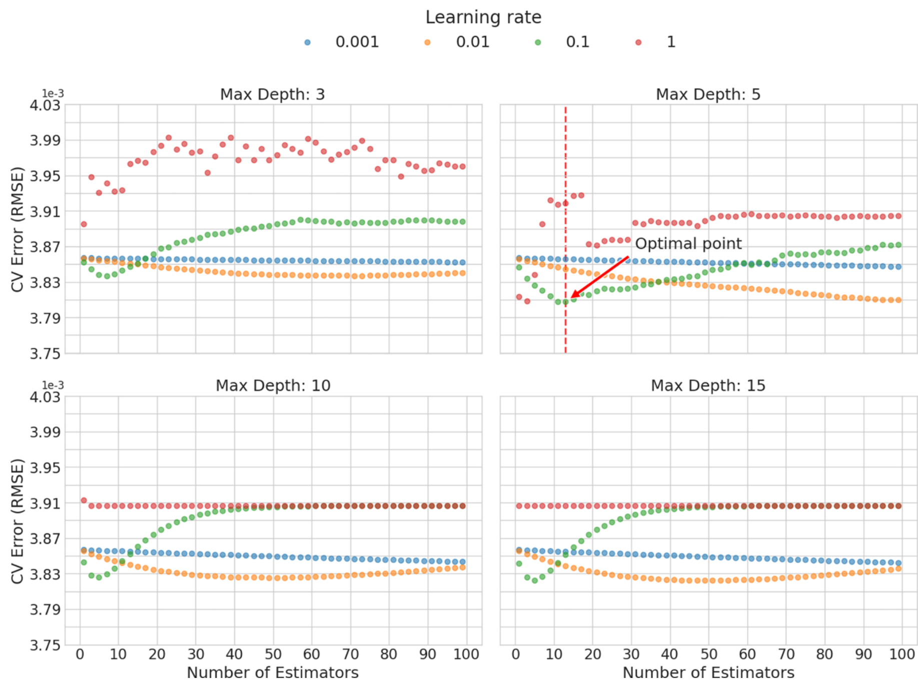

Fig. 19 Grid search results visualizing the fivefold cross-validation error across varying maximum depths, number of estimators, and learning rates in gradient boosting machine training. The dashed red line indicates the parameter combination that achieved the optimal performance

for fewer clusters to facilitate better interpretability, solidified our decision that four clusters were the most appropriate for our analysis. These clusters are referred to as spatiotemporal trip pattern passenger types: T1, T2, T3, and T4. A detailed overview of each type's key characteristics is provided in Table 1.

Appendix 5: Spatial Error Correction Model Parameter Search/Optimization

A snapshot of the grid search results is depicted in Fig. 19. The training trajectories resulting from the grid search show the cross-validation error trends across various combinations of hyperparameters for the spatial error correction model. The search explored combinations of hyperparameters including maximum depths ranging from 3 to 15, number of estimators from 1 to 100, and learning rates from 0.001 to 1. The grid size was the Cartesian product of the possible values for each hyperparameter, resulting in a total of 800 combinations evaluated using fivefold cross-validation. We identified the optimal hyperparameters as a maximum depth of 5, 13 estimators, and a learning rate of 0.1.

Acknowledgements The authors would like to thank Alex Forrest, Tolu Oke, Sandra Sheehan (Pioneer Valley Transit Authority), and Justin John (Federal Transit Administration) for supporting this research and providing the data used in the analyses. Their valuable insights and feedback contributed to the quality of this study.

Author Contributions The authors confirm their contribution to the paper as follows: study conception and design: JO; data collection: MA; algorithm development and coding: MA and JO; analysis and interpretation of results: MA and JO; draft manuscript preparation: MA and 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 used in the analyses is available at https://github.com/narslab/enhanced-trip-chaining.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

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

References

Abdalazeem M, Oke J (2023) Extracting spatiotemporal bus passenger trip typologies from noisy mobile ticketing boarding data. Data Sci Transp. https://doi.org/10.1007/s42421-023-00082-x

Alsger A, Assemi B, Mesbah M et al (2016) Validating and improving public transport origin–destination estimation algorithm using smart card fare data. Transp Res Part C Emerg Technol 68:490–506. https://doi.org/10.1016/j.trc.2016.05.004

Assemi B, Alsger A, Moghaddam M et al (2020) Improving alighting stop inference accuracy in the trip chaining method using neural networks. Public Transp 12(1):89–121. https://doi.org/10.1007/s12469-019-00218-9

Bayes T, Price M (1997) LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. Philos Trans R Soc Lond 53:370–418. https://doi.org/10.1098/rstl.1763.0053

Berkson J (1944) Application of the logistic function to bio-assay. J Am Stat Assoc 39(227):357–365. https://doi.org/10.2307/2280041

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Cui A (2006) Bus passenger origin–destination matrix estimation using automated data collection systems. Thesis, Massachusetts Institute of Technology

Currie G, Delbosc A (2011) Exploring the trip chaining behaviour of public transport users in Melbourne. Transp Policy 18(1):204–210. https://doi.org/10.1016/j.tranpol.2010.08.003

Fix E, Hodges JL (1989) Discriminatory analysis. Nonparametric discrimination: consistency properties. Int Stat Rev / Revue Internationale de Statistique 57(3):238–247. https://doi.org/10.2307/1403797

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232. https://doi.org/10.1214/aos/1013203451

Goldstein M (2016) Intelligent transportation systems: urban and rural transit providers reported benefits but face deployment challenges. Tech. rep, United States Government Accountability Office, Washington, DC

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

Gordon JB, Koutsopoulos HN, Wilson NHM et al (2013) Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. Transp Res Rec 2343(1):17–24. https://doi.org/10.3141/2343-03

Horowitz J (1980) A utility maximizing model of the demand for multi-destination non-work travel. Transp Res Part B Methodol 14(4):369–386. https://doi.org/10.1016/0191-2615(80)90017-X

Jon Scott A (1987) Long-range forecasting: from crystal ball to computer. Int J Forecast 2(3):387–390. https://doi.org/10.1016/0169-2070(86)90059-2

Kitamura R (1984) Incorporating trip chaining into analysis of destination choice. Transp Res Part B Methodol 18(1):67–81. https://doi.org/10.1016/0191-2615(84)90007-9

Lee MS, Chung JH, McNally MG (2002) An empirical investigation of the underlying behavioral processes of trip chaining. eScholarship Publishing

Lerman SR (1979) The use of disaggregate choice models in semi-markov process models of trip chaining behavior. Transp Sci 13(4):273–291. https://arxiv.org/abs/25767955

Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions and reversals. Sov Phys Dokl 10:707

Lin J (1991) Divergence measures based on the Shannon entropy. IEEE Trans Inf Theory 37(1):145–151. https://doi.org/10.1109/18.61115

Liu C, Susilo YO, Karlström A (2016) Measuring the impacts of weather variability on home-based trip chaining behaviour: a focus on spatial heterogeneity. Transportation 43(5):843–867

Marble D (1960) A theoretical exploration of individual travel behavior. In: Quantitative geography, part 1: economic and cultural topics. Northwestern University, pp 33–53

Minsky M, Papert SA (2017) Perceptrons: an introduction to computational geometry. The MIT Press. https://doi.org/10.7551/mitpress/11301.001.0001

Mohammed AA, Oke J (2023) Origin-destination inference in public transportation systems: a comprehensive review. Int J Transp 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 (2016) AWarp: fast warping distance for sparse time series. In: 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, Barcelona, Spain, pp 350–359. https://doi.org/10.1109/ICDM.2016.0046

Noland RB, Thomas JV (2007) Multivariate analysis of trip-chaining behavior. Environ Plan 34(6):953–970. https://doi.org/10.1068/b32120

O'Kelly M (2010) A model of the demand for retail facilities, incorporating multistop, multipurpose trips. Geogr Anal 13:134–148. https://doi.org/10.1111/j.1538-4632.1981.tb00721.x

Pioneer Valley Transit Authority (2023) About PVTA. http://www.pvta.com/about.php

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

Sánchez-Martínez GE (2017) Inference of public transportation trip destinations by using fare transaction and vehicle location data: dynamic programming approach. Transp Res Rec 2652(1):1–7. https://doi.org/10.3141/2652-01

Schmöcker JD, Su F, Noland RB (2010) An analysis of trip chaining among older London residents. Transportation 37(1):105–123. https://doi.org/10.1007/s11116-009-9222-z

Thill JC, Thomas I (1987) Toward conceptualizing trip-chaining behavior: a review. Geogr Anal 19(1):1–17. https://doi.org/10.1111/j.1538-4632.1987.tb00110.x

Van Der Hoorn T (1983) Experiments with an activity-based travel model. Transportation 12(1):61–77. https://doi.org/10.1007/BF00165572

Vanderwaart C (2016) Planning transit networks with origin, destination, and interchange inference. Thesis, Massachusetts Institute of Technology

Vanderwaart C, Attanucci JP, Salvucci FP (2017) Applications of inferred origins, destinations, and interchanges in bus service

planning. Transp Res Rec 2652(1):70–77. https://doi.org/10.3141/2652-08
Vintsyuk TK (1972) Speech discrimination by dynamic programming. Cybernetics 4(1):52–57. https://doi.org/10.1007/BF01074755
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
Yalamanchili L, Pendyala RM, Prabaharan N et al (1999) Analysis of global positioning system-based data collection methods for capturing multistop trip-chaining behavior. Transp Res Rec 1660(1):58–65. https://doi.org/10.3141/1660-08
Yang M, Wang W, Chen X et al (2007) Empirical analysis of commute trip chaining: case study of Shangyu, China. Transp Res Rec 2038(1):139–147. https://doi.org/10.3141/2038-18
Zadeh LA (1965) Fuzzy sets. Inf Control 8(3):338–353. https://doi.org/10.1016/S0019-9958(65)90241-X