A novel origin-destination-transfer model using mobile ticketing activations with seasonal passenger typology and spatial error correction

Mohammed Abdalazeem*[a], Tolu Oke[b], Jimi Oke[a]

[a]*Department of Civil and Environmental Engineering, University of Massachusetts Amherst, Amherst, Massachusetts, USA*
[b]*Pioneer Valley Transit Authority, Springfield, Massachusetts, USA*

Abstract

We introduce a framework to construct an origin-destination-transfer (ODX) model for regional transit agencies using mobile ticketing data. The model applies trip chaining, enhanced by spatiotemporal typology analysis and spatial error correction, to infer passenger origins, destinations, and transfers. Using the Pioneer Valley Transit Authority (PVTA) as a case study, we develop a seasonal typology-informed trip chaining framework to estimate a full network ODX matrix. We demonstrate the model's practicality through an analysis of the transit system's spatial distribution, seasonal variations in passenger flows, and major transfer points and routes. This innovative framework provides a scalable, cost-effective tool for regional transit agencies to better understand transit demand and system utilization based on a sample of mobile ticketing data.

Keywords: public transit, origin-destination matrix estimation, trip chaining, mobile ticketing data, automated fare collection, transit planning, data-driven decision-making

1. Introduction

It is important for public transit agencies to understand passenger travel patterns across the transit network in order to effectively plan service. To achieve this, agencies need to have a comprehensive picture of how passengers travel on the existing network of routes, including where riders originate, end, and transfer where applicable. Existing data sources, though effective in their capabilities to collect granular data, may not offer the level of detail required to accurately capture and understand transit travel flow patterns (Cui, 2006).

Origin-destination-transfer (ODX) models have emerged as an effective tool for transit agencies. They provide a detailed picture of where passengers board and alight, the routes taken, and transfers across the transit network. However, the traditional methods used to collect data for ODX modeling are resource-intensive and often impractical for regional transit systems (Sun and Xu, 2012).

We propose a novel and comprehensive approach for developing an origin-destination-transfer (ODX) model using mobile ticketing AFC data that builds upon our previous work

*Corresponding author: Mohammed Abdalazeem, Email: mamohammed@umass.edu

(Abdalazeem and Oke, 2023, 2024). To the best of our knowledge, this is the first study to infer origin-destination-transfer patterns directly from mobile ticketing activations. We focus on the Pioneer Valley Transit Authority (PVTA) in Massachusetts as our case study. Our approach offers a practical solution for regional transit systems facing data limitations, allowing them to develop accurate and scalable ODX models cost-effectively. By demonstrating the potential of mobile ticketing data in this context, we aim to encourage the broader adoption of similar methods across transit agencies utilizing mobile AFC systems. This, in turn, will enhance transit demand analysis, facilitate more informed and data-driven transit planning decisions, and finally contribute to the development of more efficient and responsive public transportation networks. We summarize the acronyms and mathematical notation used throughout this paper in Table 1 and Table 2, respectively.

Table 1: List of acronyms used in this paper.

Acronym	**Description**
AFC	Automated Fare Collection
APC	Automated Passenger Counter
AVL	Automated Vehicle Location
GBM	Gradient Boosting Machine
GTFS	General Transit Feed Specification
IPF	Iterative Proportional Fitting
JSD	Jensen-Shannon Divergence
OD	Origin-Destination
ODX	Origin-Destination-Transfer
PVTA	Pioneer Valley Transit Authority

2. Literature review

Origin-destination (OD) modeling is fundamental for planning and operational decisions in public transportation. Early approaches relied on manual surveys to estimate OD flows, as exemplified by pioneering works in the 1970s and 1980s (Wilson, 1970; Robillard, 1975; Ben-Akiva et al., 1985). However, the advent of automated data collection and mobile communication technologies revolutionized the field, enabling high-resolution data collection on passenger movements (Zannat and Choudhury, 2019; Mohammed and Oke, 2023).

Recent OD modeling has leveraged large-scale data sources, primarily passenger-linked smart card data, to estimate route-level and full network flows in transit systems. In addition, automated fare collection (AFC), automated passenger counter (APC), and automated vehicle location (AVL) systems have become prevalent, offering a more detailed alternative to traditional methods (Cui, 2006; Nassir et al., 2011). Studies like Cui (2006) and Gordon (2012) have demonstrated the potential of AFC and smart card data to enhance OD flow estimation and analyze passenger behavior, respectively.

The focus has shifted towards practical implementations, with an emphasis on trip chaining to understand and improve the relationship between OD estimation and transport planning (Alsger et al., 2015; Fan and Chen, 2018; Huang et al., 2020; Gordon, 2012). Additionally, machine learning and artificial intelligence have been integrated to re-imagine traditional OD estimation approaches (Cheng et al., 2019; Nassir et al., 2011; Sánchez-Martínez,

Table 2: Mathematical notation used in this paper.

Notation	Description
$\mathbf{A}$	Set of mobile ticketing activation
$\hat{F}_{ij}$	Final scaled flow between origin i and destination j from the Iterative Proportional Fitting (IPF) process
$\hat{F}_{ijk}$	Final scaled flow between origin i and destination j with transfer at stop k from the Iterative Proportional Fitting (IPF) process
$\hat{f}_{ij}$	Inferred flow of passengers between origin i and destination j
$\hat{f}_{ijk}$	Inferred flow of passengers who transfer at stop k when traveling from origin i to destination j
$\mathcal{G}(\cdot)$	Gradient boosting machine model
$\mathbb{I}(\cdot)$	Indicator function
m^*	Optimal season length
$\hat{N}^z_{x,i}$	Inferred number of passengers performing activity z from route x at stop i
$\hat{N}^z_i$	Inferred total number of passengers performing activity z at stop i
p_{abc}	Observed survey frequency for zones a, b, and c
$\hat{p}_{abc}$	Predicted frequency for zones a, b, and c
$\tilde{p}_{abc}$	Corrected frequency for zones a, b, and c
$\hat{p}_{ij}$	Estimated flow frequency for origin-destination (OD) pair (i, j)
$\hat{p}_{ijk}$	Estimated probability of transferring at stop k for trips from origin i to destination j
$\hat{\mathbf{R}}$	Set of chained trips with inferred stops and a binary transfer indicator
$\mathbf{r}$	Vector of trip chain information, including origin r_o, destination r_d, and transfer stop (if any) r_t
$\mathbb{TC}(\cdot)$	Trip chaining algorithm
z	An activity type (boarding, alighting, or transferring)
ϵ_{abc}	Residual between observed and predicted frequencies for zones a, b, and c
θ^*_m	Optimal parameter vector for seasonal period (m)
κ^*_m	Optimal distance threshold quantile
$\nu^z_{x,i}$	Route usage frequency for route x at stop i for activity z
$\rho_{A,c,t}$	Seasonal type-specific distance threshold for alighting
$\rho_{B,c,t}$	Seasonal type-specific distance threshold for boarding
$\rho_{T,c,t}$	Seasonal type-specific distance threshold for transfers
τ^*_m	Optimal transfer time threshold

2017). These methods often rely on comprehensive smart card-based AFC data, limiting their applicability to small or mid-sized transit agencies.

The rise of mobile ticketing during the COVID-19 pandemic among small- and mid-sized transit agencies, many of which previously lacked passenger-linked travel data, presented an opportunity to access detailed data for ODX modeling. However, there was a gap in the ODX modeling approaches for effectively utilizing the often lower quality mobile ticketing data sources with advanced methods to achieve the same level of detail as smart card-based approaches.

Further, despite advancements in understanding passenger flow dynamics, challenges remain in capturing the dynamic nature of urban mobility and transfer behavior in the modeling approaches (Hussain et al., 2021). The integration of emerging data sources like mobile ticketing remains under-explored, and existing models often struggle to adapt to real-time changes.

Our research pioneers the integration of mobile ticketing data into OD matrix modeling,

expanding the data foundation and capturing the dynamic nature of urban transit usage. We develop a replicable framework that addresses the unique complexities of bus inference from a sample of linked mobile ticketing data, while also leveraging it for trip chaining and expansion to the full transit network population acquired from the unlinked APC boarding and alighting data. This approach is validated through survey data to ensure accuracy and reliability.

3. Methodology

The ODX model was constructed by first performing passenger typology analysis to identify distinct travel behavior groups. Next, a trip chaining framework that integrates insights from the passenger typology analysis and accounts for seasonal variations was developed to obtain a seed OD matrix that captures ridership frequencies. A spatial error correction followed to enhance the seed matrix by accounting for insufficient data collection and small sample sizes. Results from this process were then scaled to estimate a full network ODX matrix. Figure 1 shows the key steps of this process.

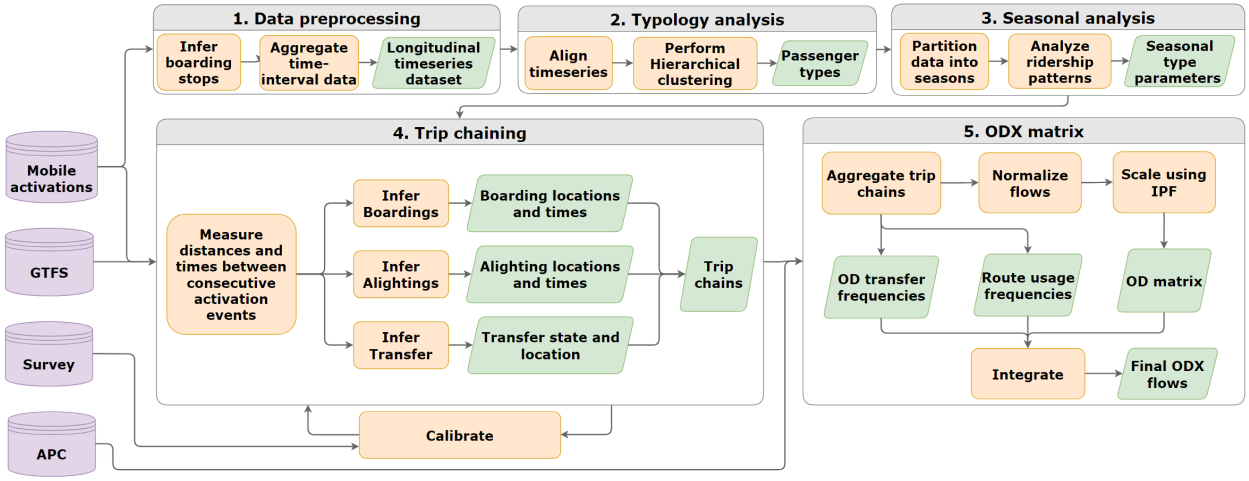


Figure 1: Flowchart of the origin-destination-transfer (ODX) modeling framework. Purple containers indicate datasets; orange boxes indicate procedures; while green parallelograms indicate intermediate/final outputs.

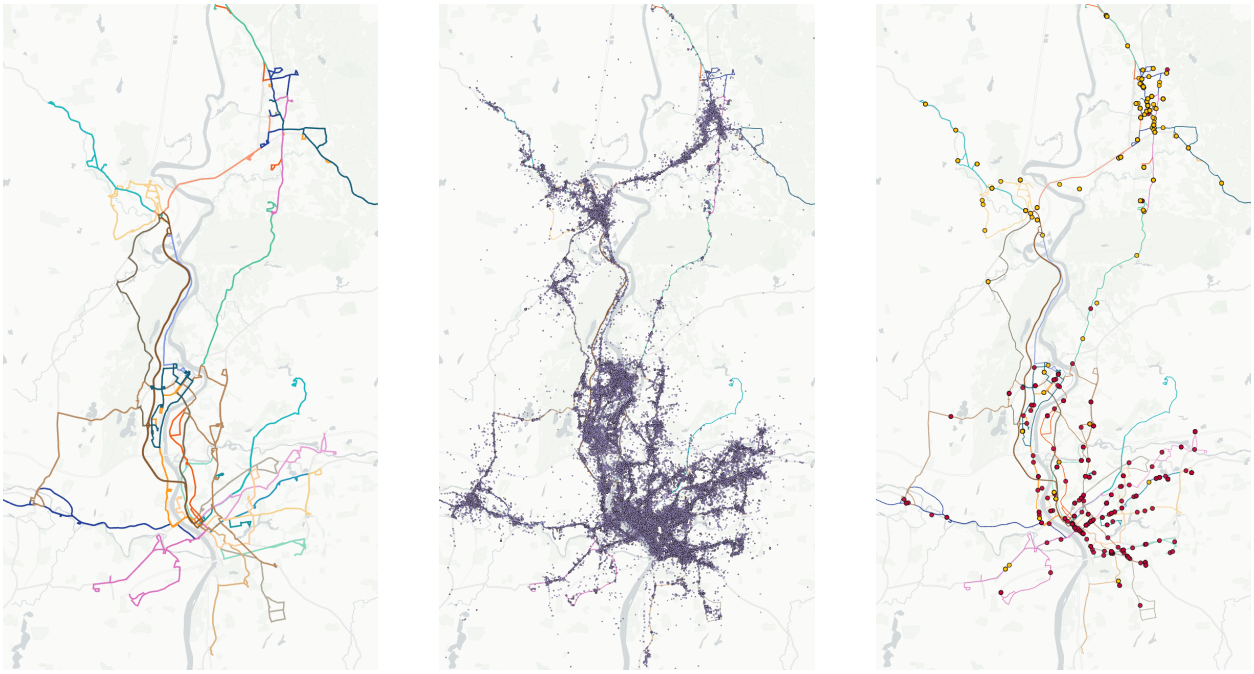
3.1. Data

The PVTA, the largest regional transit authority in Massachusetts, operates 36 fixed bus routes over a 600-square-mile area (Pioneer Valley Transit Authority, 2023). The model inputs obtained from PVTA include a sample of raw linked mobile ticketing data, full population unlinked Automated Passenger Counter (APC) data, GTFS bus schedule data, and passenger travel surveys. The bus network is shown in Figure 2(a).

The mobile ticketing data used for the trip chaining spanned from July 2020, when the mobile ticketing application was launched, to December 2022, when the data were provided for the study. It included approximately 980,000 activations from around 14,500 unique passengers, which represented a 16% sample of all fare-paying rides. The ticketing system

recorded the ticket type, activation timestamp, and geolocation (when available) upon user activation of an app-based mobile ticket Figure 2(b) shows the spatial distributions of the mobile ticket activations, while Figure 3 provides a visual analysis of the AFC data.

Since the mobile tickets are activated at or around bus stops, they first need to be linked to a specific bus stop, bus route, and bus trip for the modeling process. We performed this linkage using GTFS data, which included schedules, stop locations, and route information. This helped to identify boarding locations and potential transfers to ensure accurate modeling. We also used customer travel surveys from prior years conducted by the PVTA to calibrate and validate the model. The 2019 and 2022 PVTA travel surveys provided detailed passenger reports of prior trips' origin, transfer, and destination bus stops, and were used as ground-truth observations for the trip chaining model (Figure 2(c)). Finally, we used APC data, obtained from sensors on buses, to expand the mobile ticketing results to the full network ridership model. The APC data contained passenger counts of boarding and alighting at each bus stop (Figure 4).



(a) (b) (c)

Figure 2: PVTA network and mobile activations: (a) Map of network routes, (b) Distribution of user activations from July 2020 to December 2022, and (c) Survey data locations as conducted in 2019 (red) and 2022 (yellow).

3.2. Passenger typology analysis

We identified distinct passenger types based on their travel patterns to develop a more accurate ODX model (Abdalazeem and Oke, 2023). First, we used fast dynamic time warping (Salvador and Chan, 2004) to measure the dissimilarity between passenger trajectories extracted from the mobile ticketing data in both spatial and temporal dimensions. Second, we clustered passengers using hierarchical clustering based on the similarity of their travel

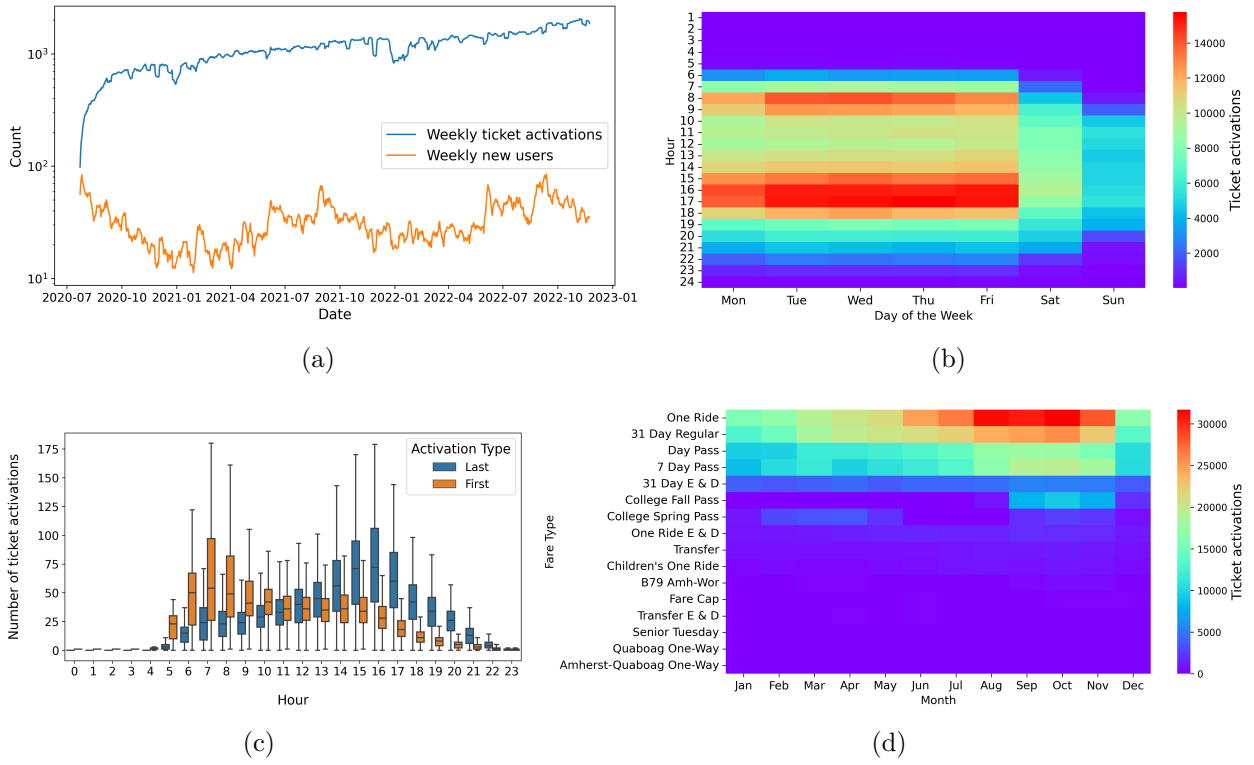


Figure 3: Overview of mobile ticketing from July 2020 through December 2022: (a) Weekly ticket activations and new users, (b) Hourly ticket activations by day of week, (c) Distribution of the first and last activations of the day, and (d) Trends of the monthly faretype usage.

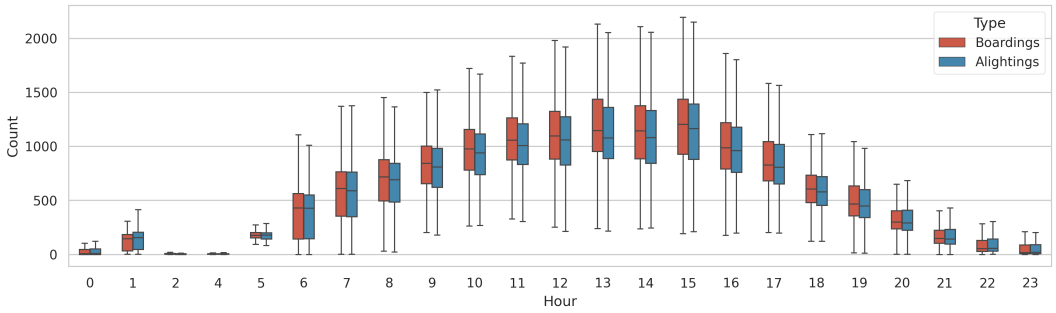
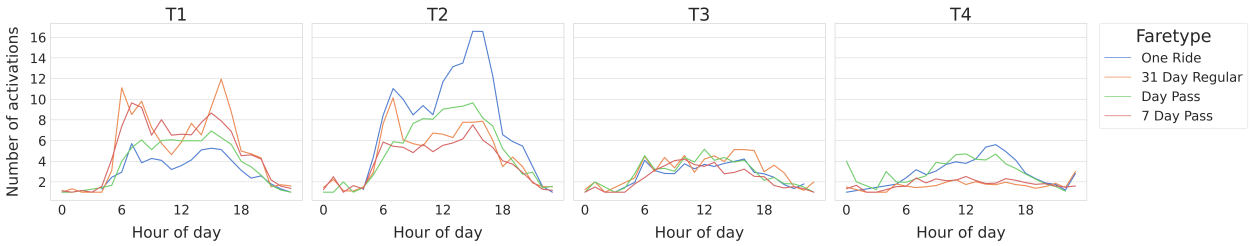


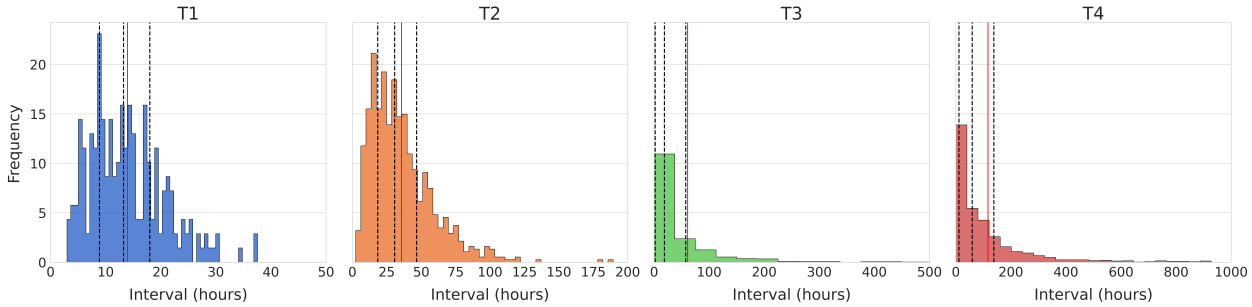
Figure 4: Hourly counts of Automated Passenger Counter (APC) boardings and alightings for the entire network.

behaviors. Observing the metrics like silhouette and within-cluster sum of squares, we found that four clusters were the optimal number to represent distinct passenger types. Figure 5 shows a visualization of the varying ridership patterns of the discovered types.

The insights gained from this typology analysis (Abdalazeem and Oke, 2023) were incorporated into the subsequent stages of the ODX model development. By accounting for the unique characteristics of each passenger type, we aimed to extract type-specific parameters that enhance the model's accuracy and provide a more nuanced understanding of the passenger network utilization patterns across the network.



(a)



(b)

Figure 5: Passenger type patterns: (a) Hourly user activations by the top used faretypes for each passenger type, and (b) Distribution of passenger time interval between activations; vertical dashed lines indicate first, second, and third quartiles, while the solid red line represents the mean.

3.3. Trip chaining

The trip chaining framework inferred alighting, boarding and transfer activities for each passenger, pinpointing the precise bus stop locations where passengers began and ended their journeys (Abdalazeem and Oke, 2024). A binary transfer indicator was also determined for each trip. For trips with a transfer, the precise transfer bus stop was inferred by combining consecutive trip chains for the same user. The origin and destination of the combined trip were designated as the origin of the first trip chain and the destination of the second trip chain, respectively, and the transfer stop was designated as the origin of the second trip chain. By following this process, the framework reconstructed complete trip chains, representing distinct journeys taken by passengers for various purposes. The framework is given by:

$$\hat{\mathbf{R}} = \mathbb{TC}(\mathbf{A}; \boldsymbol{\theta}_{m^*}) \tag{1}$$

where $\mathbb{TC}$ represents the trip chaining algorithm, including boarding ($\mathbb{BI}$), alighting ($\mathbb{AI}$), and transfer inferences ($\mathbb{TI}$). $\mathbf{A}$ represents the set of activations, and $\hat{\mathbf{R}}$ is the resulting chained trips with each user trip chain $\mathbf{r} \in \hat{\mathbf{R}}$ representing a vector of inferred trip chain information, including origin r_o, destination r_d, and transfer stop (if any) r_t. The parameter vector $\boldsymbol{\theta}_{m^*} = [\kappa^*_{m^*}, \tau^*_{m^*}]$ includes the optimal distance threshold quantile $\kappa^*_{m^*}$ and the optimal transfer time threshold $\tau^*_{m^*}$. The distance threshold is used to determine type-specific distances for boarding ($\rho_{B,c,t}$), alighting ($\rho_{A,c,t}$), and transfers ($\rho_{T,c,t}$), while the transfer time threshold defines the maximum time between events for a transfer to be valid. Both parameters were optimized against survey data to ensure the framework's accuracy, and the optimal season length m^* was chosen to minimize the divergence between predicted and observed trip frequencies, as discussed in §3.3.1.

3.3.1. Parameter optimization

We aggregated the stop-level mobile ticketing activation data into geographically defined zones, as shown in Figure 6, to match the resolution of the passenger survey data and simplify parameter optimization. This aggregation enabled a direct comparison between the model's predictions and the survey observations, which allowed us to fine-tune the parameters for accuracy.

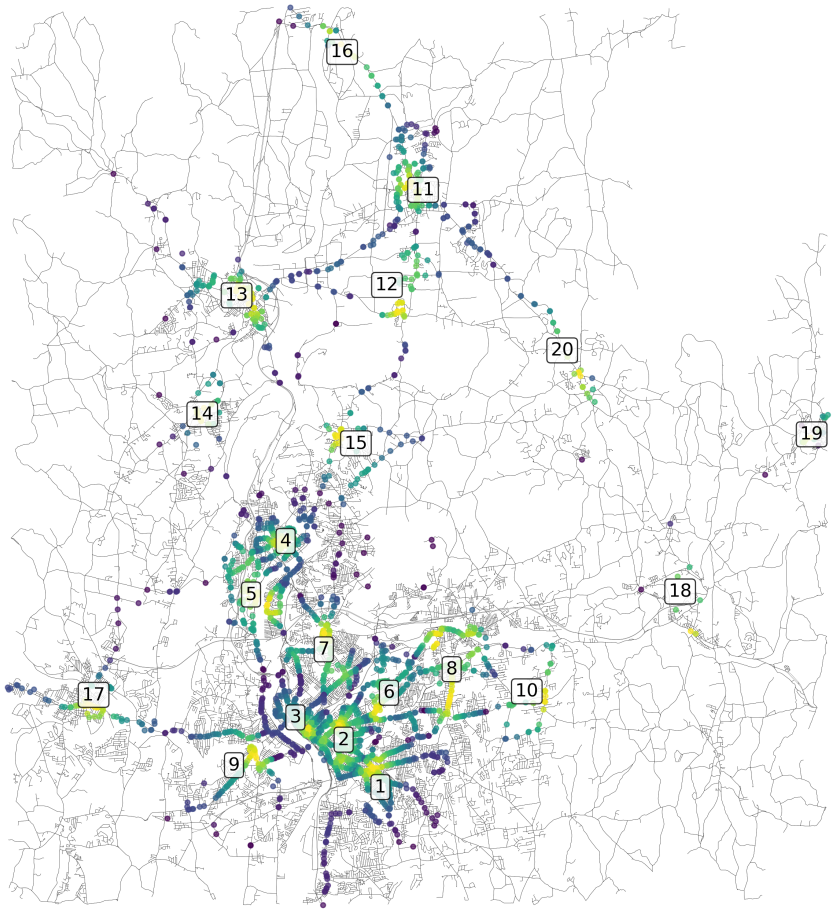


Figure 6: Geographic zones used for data aggregation and model validation.

Seasonal variations can play an important role in public transit systems, especially in regions with diverse weather conditions and fluctuating activity patterns across the year. For example, changes in the academic calendar, holiday travel, and weather conditions significantly affect passenger flow and transfer patterns. To account for these variations, we partitioned the mobile ticketing activation data into multiple periods of varying lengths ($m \in \{0, 1, 2, 3, 4\}$ months) and performed a grid search to optimize the seasonal period for segmenting the data, along with different values for the distance threshold quantile (κ_m) and transfer time threshold (τ_m). The optimal parameter combination was selected by minimizing the Jensen-Shannon Divergence (JSD) between predicted and observed trip frequencies from the survey data.

The error trajectories for different combinations of transfer time thresholds (τ_{m^*}) and quantiles (κ_{m^*}) are shown in Figure 7. The optimal point was determined with a seasonal period of $m^* = 4$ months, $\kappa_{m^*} = 0.7$, and $\tau_{m^*} = 35$. Consequently, we divided the year into distinct seasonal periods: Spring Semester (January–April), Summer Break (May–August), and Fall Semester (September–December). This segmentation allowed us to capture how ridership and transfer behavior fluctuate throughout the year. Finally, the optimized parameters were used in the trip chaining framework to estimate complete trip chains from the mobile ticketing data (Abdalazeem and Oke, 2024).

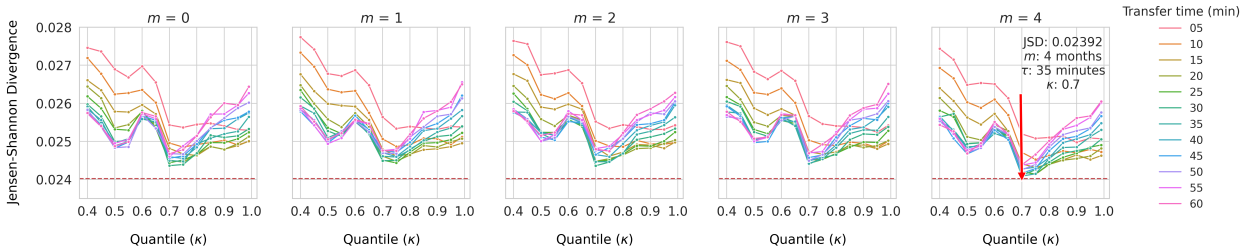


Figure 7: Error trajectories for different combinations of transfer time threshold (τ_{m^*}) and quantile (κ_{m^*}), showing optimal settings that minimize error metrics.

3.3.2. Seed matrix construction

As defined in Equation 1, the trip chaining framework produced a set of chained trips $\hat{\mathbf{R}}$. We aggregated the activity information within $\hat{\mathbf{R}}$ to calculate the inferred flow of passengers $\hat{f}_{ij}$ between origin i and destination j, and the flow of passengers $\hat{f}_{ijk}$ who transfer at stop k when traveling from origin i to destination j. To do this, we counted the occurrences of each origin-destination and origin-destination-transfer combination in the set of chained trips $\hat{\mathbf{R}}$.

$$\hat{f}_{ij} = \sum_{\mathbf{r} \in \hat{\mathbf{R}}} \mathbb{I}(r_o = i, r_d = j) \tag{2}$$

$$\hat{f}_{ijk} = \sum_{\mathbf{r} \in \hat{\mathbf{R}}} \mathbb{I}(r_o = i, r_d = j, r_t = k) \tag{3}$$

where $\mathbb{I}(\cdot)$ is an indicator function that equals 1 if the condition $(\cdot)$ is true and 0 otherwise, and r_o, r_d, and r_t are the origin, destination, and transfer stop (if any) of trip chain $\mathbf{r}$, respectively.

We then constructed a high-resolution stop-level seed OD matrix using the calibrated model. The seed matrix captured the initial estimates of passenger flows between origins and destinations, ignoring transfers. The flow frequency $\hat{p}_{ij}$ for each OD pair (i, j) was calculated as:

$$\hat{p}_{ij} = \frac{\hat{f}_{ij}}{\sum_i \sum_j \hat{f}_{ij}} \tag{4}$$

In addition to the seed matrix, we calculated transfer probabilities. For each OD pair, we identified potential transfer stops k and calculated the probability of passengers transferring at each stop. This was done by dividing the number of passengers observed transferring at

a particular stop by the total number of passengers traveling between that OD pair:

$$\hat{p}_{ijk} = \frac{\hat{f}_{ijk}}{\sum_k \hat{f}_{ijk}} \tag{5}$$

where $\hat{p}_{ijk}$ is the estimated probability of transferring at stop k for trips from origin i to destination j.

3.3.3. Spatial error correction

We incorporated a spatial error correction model to address limitations in mobile ticketing implementation. This approach allowed us to estimate travel patterns in areas where mobile ticketing data may be incomplete or unavailable, such as regions with free fares or low mobile app usage.

To implement this model, we first aggregated the seed matrix to the zonal level, aligning it with the resolution of the survey data. We then used a gradient boosting machine (GBM) model $\mathcal{G}$, trained on a year of AFC data, to learn and correct spatial errors in the zonal trip chaining predictions. We performed a grid search to optimize the GBM model's hyperparameters, including the number of trees, learning rate, maximum depth, and minimum samples split. The optimal hyperparameters were selected based on minimizing the mean squared error (MSE) on a validation set.

The model used the residuals ϵ_{abc}, defined as the difference between observed survey fequencies p_{abc} and predicted frequencies $\hat{p}_{abc}$ for zones a, b, and c:

$$\epsilon_{abc} = p_{abc} - \hat{p}_{abc} \tag{6}$$

The corrected frequency $\tilde{p}_{abc}$ is given by:

$$\tilde{p}_{abc} = \hat{p}_{abc} + \hat{\epsilon}_{abc} = \hat{p}_{abc} + \mathcal{G}(a, b, c) \tag{7}$$

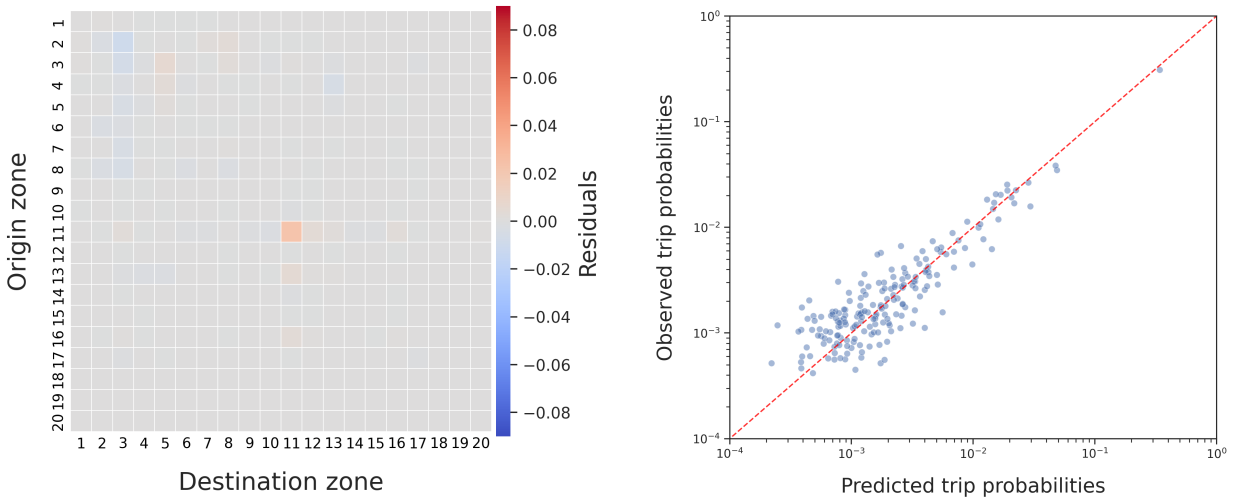
where $\mathcal{G}(a, b, c)$ is the GBM model's prediction for the residual for the given zone combination. We then disaggregated the resulting zonal seed matrix back to the stop level, maintaining the original proportions of the seed matrix. This approach ensured that the enhanced trip chaining predictions are available at the stop level for subsequent analysis and modeling steps. We validated the results using five months of test data (from August through December 2021).

3.3.4. Performance

We compared the inferred trip frequencies with those observed in the survey data to assess the performance of the trip chaining framework. Figure 8 presents a heatmap of the residuals between predicted and observed trip frequencies and a scatter plot of observed versus predicted trip probabilities, which demonstrate the model's accuracy.

3.4. Estimation of the full network ODX matrix

After reconstructing individual passenger journeys through trip chaining, we built a comprehensive representation of passenger flows across the entire transit network, capturing transfers between routes. Mobile ticketing data generally lacks transfer totals and detailed boarding/alighting information. To address this, we implemented a two-step process. First, we scaled the initial origin-destination (OD) matrix to match the Automated Passenger



(a) (b)

Figure 8: Trip chaining model validation: (a) Heatmap of residuals between predicted and observed trip frequencies and (b) Scatterplot of observed versus predicted trip probabilities, demonstrating the model's accuracy.

Counter (APC) ridership data using Iterative Proportional Fitting (IPF). Second, we used the estimated transfer probabilities from the trip chaining results to transform the two-dimensional OD matrix into a three-dimensional ODX matrix and capture passenger transfer flows in detail.

3.4.1. Iterative proportional fitting

Iterative proportional fitting (IPF) adjusts the seed matrix derived from mobile ticketing data to align its row and column sums with known boarding and alighting counts obtained from the APC system. This iterative process preserves the relative distribution of flows within the matrix while ensuring it accurately reflects observed ridership totals. The IPF update is given by:

$$F_{ij}^{(r+1)} = F_{ij}^{(r)} \times \frac{O_i}{\sum_n F_{ij}^{(r)}} \quad \text{(row adjustment)} \tag{8}$$

$$F_{ij}^{(r+2)} = F_{ij}^{(r+1)} \times \frac{D_j}{\sum_m F_{ij}^{(r+1)}} \quad \text{(column adjustment)} \tag{9}$$

where $F_{ij}^{(r)}$ is the estimated flow between origin i and destination j at iteration r, O_i is the observed boarding total at origin i from APC data, and D_j is the observed alighting total at destination j from the APC data. This process iterates, adjusting rows and columns until convergence is reached. Figure 9 shows the convergence plot of the IPF process on our seed matrix.

3.4.2. Expanding to three-dimensional ODX matrix

We incorporated transfer flows into the scaled two-dimensional OD matrix (resulting from the IPF process described in §3.4.1) using the transfer probabilities $\hat{p}_{ijk}$ from §3.3.2. These

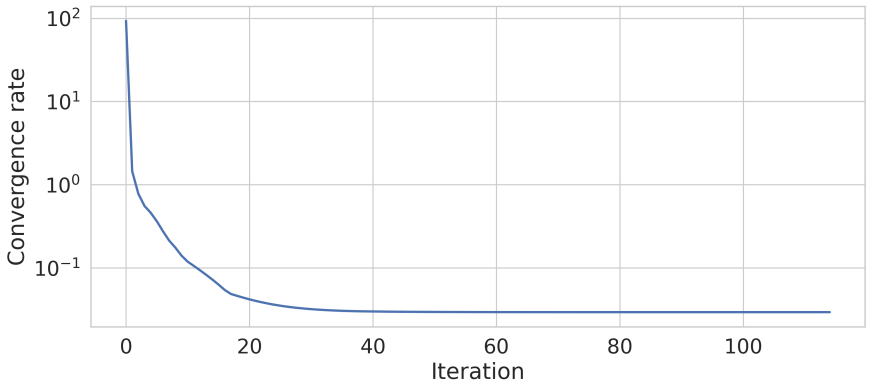


Figure 9: Convergence of the Iterative Proportional Fitting (IPF) algorithm.

probabilities represent the likelihood of a passenger transferring at stop k when traveling from origin i to destination j. The flow $\hat{F}_{ij}$ between each OD pair was distributed across potential transfer stops k based on these probabilities:

$$\hat{F}_{ijk} = \hat{p}_{ijk}\hat{F}_{ij} \tag{10}$$

where $\hat{F}_{ij}$ is the final scaled flow from the IPF process. The original OD flow was then updated to reflect passengers who did not transfer:

$$\hat{F}'_{ij} = \hat{F}_{ij} - \sum_{k} \hat{F}_{ijk} \tag{11}$$

This resulted in a three-dimensional ODX matrix capturing both direct and transfer trips.

3.4.3. Estimating route usage frequency

Finally, we estimated the frequency with which specific routes are used at each stop, considering boarding, alighting, and transfer activities. Leveraging the APC and AFC data, we determined the route usage frequency $\nu^z_{x,i}$ for each stop i, route x, and activity z (boarding, alighting, or transferring):

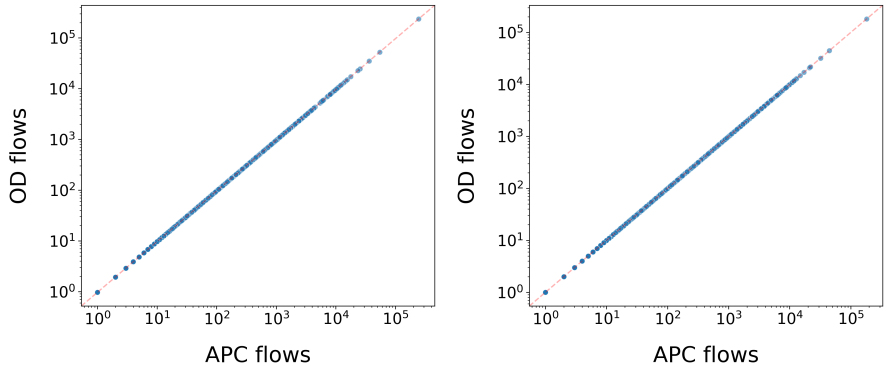
$$\nu^z_{x,i} = \begin{cases} \frac{\hat{N}^z_{x,i}}{\hat{N}^z_i} & \text{if } z = t \\ \frac{N^z_{x,i}}{N^z_i} & \text{if } z \neq t \end{cases} \tag{12}$$

where $N^z_{x,i}$ and $\hat{N}^z_{x,i}$ represent the actual and inferred number of passengers performing activity z from route x at stop i, respectively, while N^z_i and $\hat{N}^z_i$ represent the actual and inferred total number of passengers performing activity z at stop i, respectively. For transfers, we rely on the inferred total for transfers since the actual total number of passengers performing a transfer is not available in the APC data.

These estimates are limited to mobile ticketing users. However, they offer valuable insights for transit agencies to identify popular routes and stops, optimize service frequencies, and inform infrastructure improvements.

4. Results and discussion

We validated the ODX model against Automated Passenger Counter (APC) data and found a strong correlation between the two datasets, as shown in Figure 10. The model accurately captured overall passenger flows, with all points aligning closely with the 45-degree line, indicating near-perfect agreement. The Pearson correlation coefficient between the ODX model and APC data was 0.98, which confirms the model's ability to reliably estimate transit ridership patterns.



(a) (b)

Figure 10: origin-destination-transfer (ODX) totals vs. Automated Passenger Counter (APC) totals with a line of perfect agreement for (a) Boardings and (b) Alightings.

4.1. ODX flow analysis

The final three-dimensional ODX matrices capture 1,069 origin stops, 1,259 destination stops, and 502 transfer stops. They show a total inferred flow of 31,918,000 rides in 2022, with approximately 4,139,000 transfers, accounting for 13% of all trips. Figure 11 shows the distribution of flows across all origin-destination-transfer (ODX) combinations, with an average flow of 258 rides per OD pair and a standard deviation of 1,215.

Figure 12 presents heatmaps showing the total seasonal passenger flows between zones for each season. The variations in these flows reveal the changes in ridership patterns throughout the seasons. Key statistics from the ODX matrix, summarizing total trips, average trip length, and percentage of transfers for each season in 2022, are presented in Table 3. Season 2 (Summer Break) exhibited the highest total flow and the highest percentage of transfers.

Table 3: Summary of key statistics from the ODX matrix.

Season	**Total flow (10^7)**	**Avg. trip length (km)**	**Percentage of transfers**
Spring Semester	1.08	3.0	10.21
Summer Break	1.13	2.9	19.02
Fall Semester	0.97	3.0	8.98

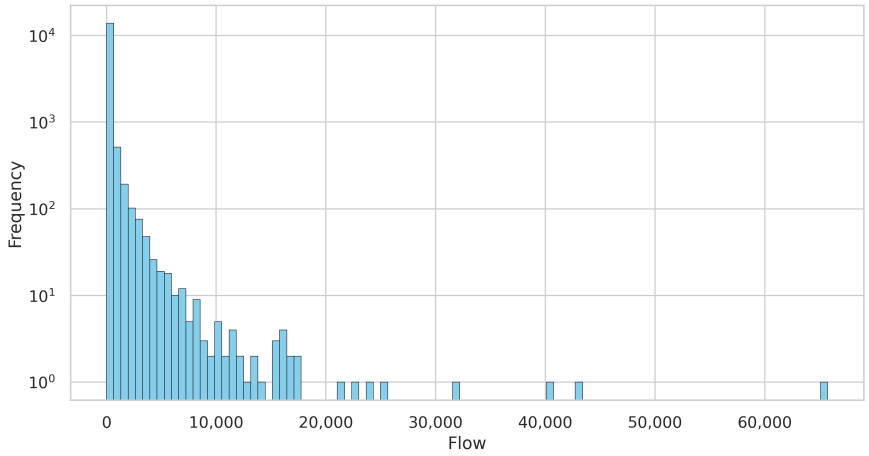


Figure 11: Distribution of inferred transit flows for all origin-destination-transfer (ODX) combinations.

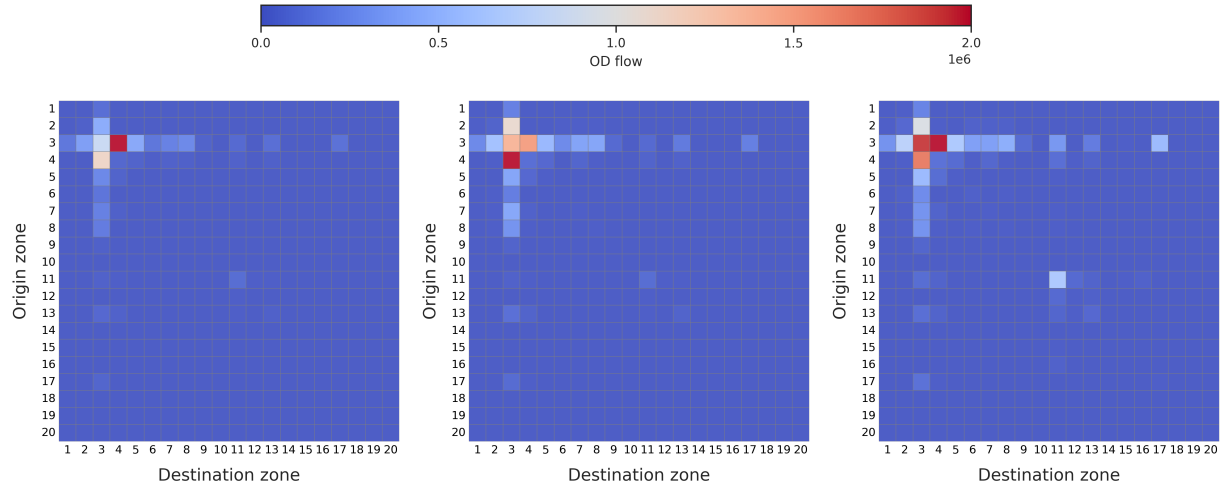


Figure 12: Heatmaps of total passenger flows between zones showing the seasonal variation in ridership for: (a) Spring Semester, (b) Summer Break, and (c) Fall Semester.

4.2. Transfer analysis

To better understand the dynamics of transit usage, we examined the ODX flows across the distinct zones and seasons in 2022. As shown in Figure 13, there are significant variations in transfer proportions across both time and location. For example, Wilbraham (Zone 10), a suburban town, experienced a very high proportion of transfers in the Spring Semester (Season 1), contrasting with the minimal activity observed in the Summer Break (Season 2) and Fall Semester (Season 3). Conversely, Springfield (Zone 2), the main urban center in the region, had a high proportion of transfers in the Fall Semester, with relatively lower activity in the other seasons.

These observed trends could be attributed to several factors, including increased winter ridership in Wilbraham for commuting and accessing essential services in Springfield, potentially due to inclement weather. Conversely, the lower transfer activity in the spring and

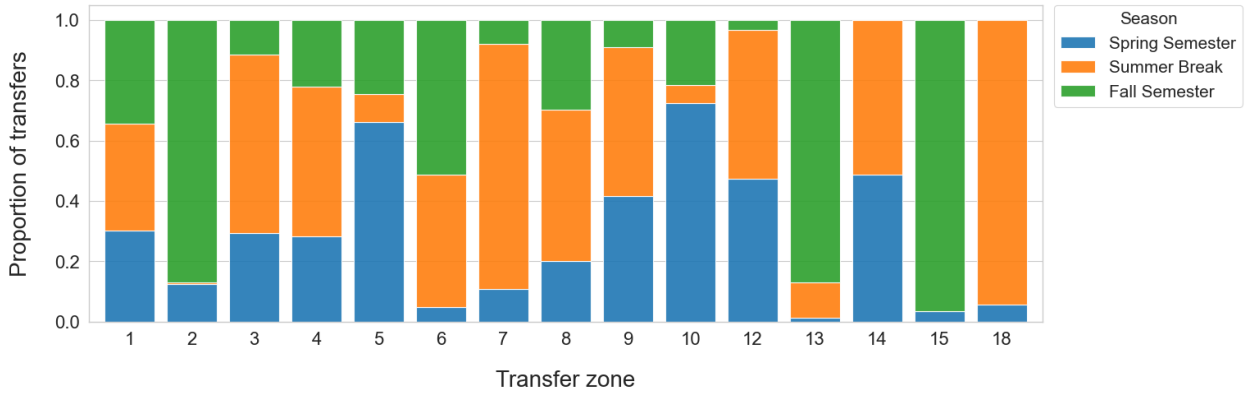


Figure 13: Proportion of transfers across different zones (x-axis) during the three distinct seasons in 2022. Zones not shown had no transfers during this year.

summer could be due to increased personal vehicle use for recreational activities. In Springfield, the higher proportion of fall transfers may be linked to the start of the academic year, while lower transfer activity in other seasons could be due to school breaks and holidays. Further investigation is needed to fully understand these seasonal and spatial variations.

These findings offer valuable insights for transit planners and operators. For instance, high-demand routes might benefit from increased service frequency or larger vehicles, while routes with more variable demand might require flexible scheduling or targeted marketing. Understanding these nuanced patterns allows for better resource allocation and ultimately improved service quality.

4.3. Insights and applications

The ODX model provides valuable insights into passenger travel behavior within the network, offering actionable information for improving service planning and operational efficiency, especially in the context of regional and rural bus networks. Unlike larger urban transit systems, these networks often face unique challenges, including lower ridership densities, longer route distances, and a higher reliance on transfers. The ODX model offers a data-driven approach to address these challenges by providing a granular understanding of passenger flows, transfer patterns, and route usage.

A key application from the ODX model is the identification of major transfer points within the network. Figure 14(a) shows the spatial distribution of these transfer points. For instance, the model reveals that Holyoke Transportation Center (HTC) and Springfield Union Station are key bus transfer hubs. This information is useful for optimizing transfer infrastructure, such as improving waiting areas, signage, and schedule coordination, to enhance passenger experience and reduce transfer times.

Another application of the ODX model is to identify high-transfer routes, which is valuable for regional and rural bus networks where transfers are often unavoidable due to longer route distances and lower ridership. Figure 14(b) shows some of the routes predominantly used for transfers between popular OD pairs. This information allows transit planners to assess the efficiency of existing transfer connections and identify opportunities for improvement. For instance, introducing a direct service or adjusting existing schedules to minimize transfer times could significantly enhance passenger experience.

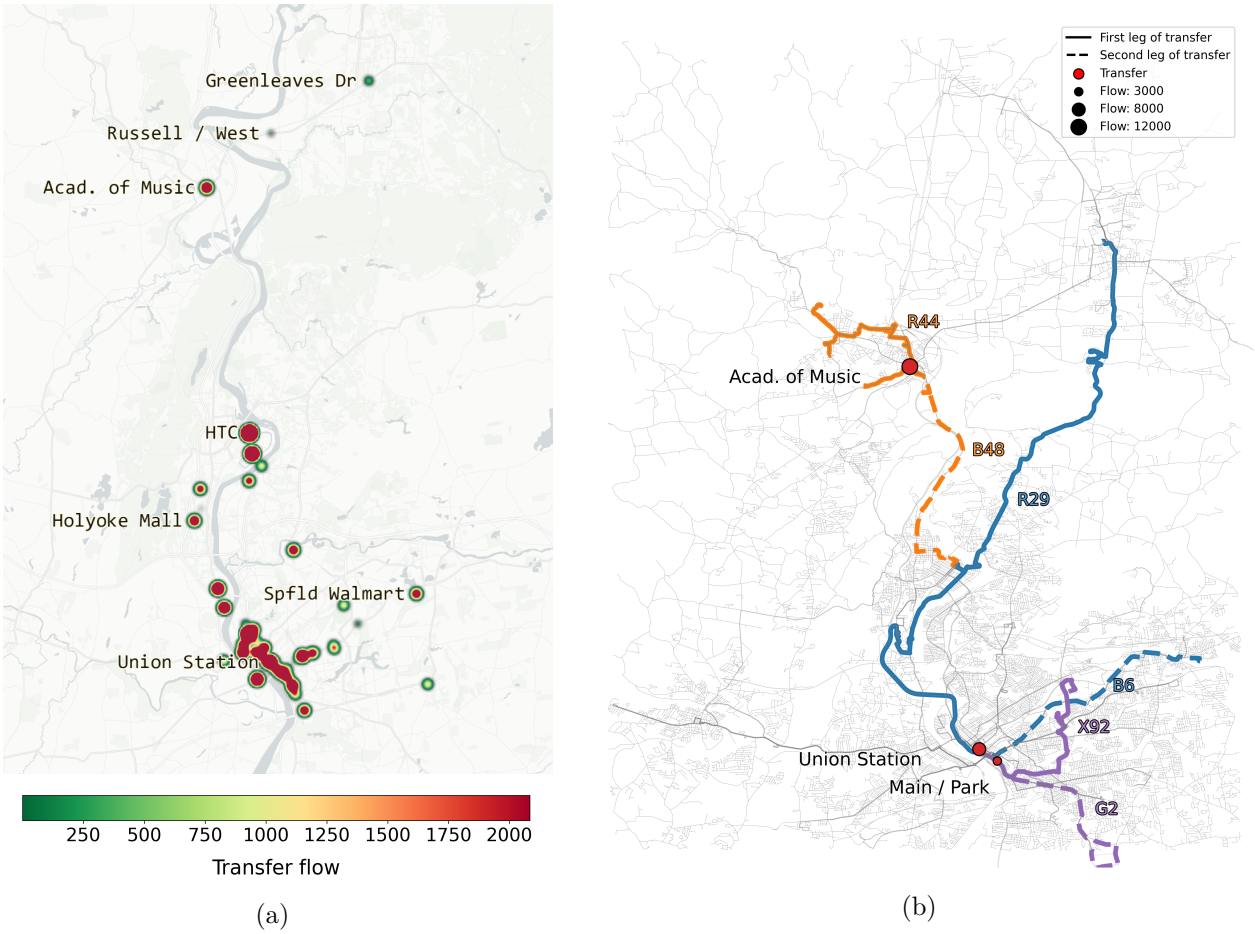


Figure 14: Key transfer locations and routes: (a) Areas with a high frequency of transfers and (b) Selected routes used to transfer between popular origin-destination (OD) pairs that require a transfer.

Beyond the identification of key transfer hubs and routes, the ODX model also provides a better understanding of passenger movement by detecting the most frequently used OD pairs that involve transfers. This level of detail allows transit agencies to make more informed and targeted service improvements but is often missed in traditional transit data. Figure 15 shows the top 10 OD pairs with the highest transfer flow, which highlights the busy transfer points as well as the specific routes passengers are using to connect from these points to their origins and destinations.

This information allows transit agencies to make strategic decisions, such as optimizing connections between key routes or introducing direct services to reduce the need for transfers, especially for high-demand OD pairs. By making these targeted investments, transit planners can enhance passenger experience, minimize travel times, and boost overall system efficiency. These insights from the ODX framework provide a level of precision that traditional methods often struggle to achieve, which helps in improving network connectivity.

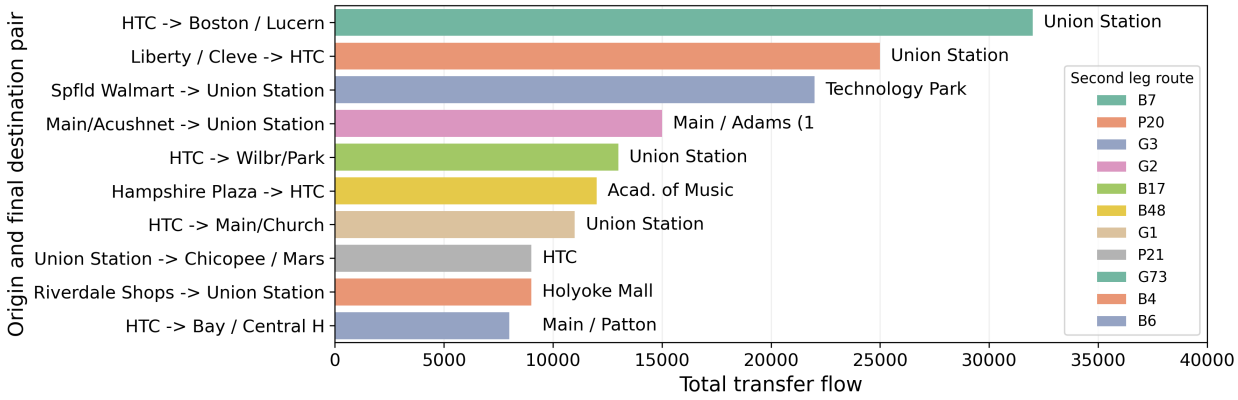


Figure 15: Top 10 origin-destination (OD) pairs with the highest transfer flow and their associated routes. The color of each bar indicates the second leg of the journey (the route taken after the transfer), and the label in front of each bar identifies the specific transfer point where passengers switch between routes.

5. Conclusion

In this paper, we presented a comprehensive framework for constructing an origin-destination-transfer (ODX) model using mobile ticketing data, specifically tailored for regional transit agencies like the Pioneer Valley Transit Authority (PVTA). By leveraging a sample of mobile ticketing activations, we developed a trip chaining method enhanced with passenger typology analysis and seasonal variations to accurately infer detailed passenger travel patterns, including origins, destinations, and transfers. The inclusion of spatial error correction and the application of Iterative Proportional Fitting (IPF) further refined the model, allowing us to estimate a full network ODX matrix that closely aligns with actual ridership patterns, as validated against survey and Automated Passenger Counter (APC) data.

Our case study of the PVTA network demonstrated the ODX framework's effectiveness in capturing complex transit dynamics in a regional context. The model successfully identified key transfer points, seasonal variations in passenger flows, and high-transfer routes, providing valuable insights for transit planners and operators. These insights can inform service optimization, infrastructure improvements, and strategic decision-making, especially in regional and rural bus networks where data limitations have historically hindered detailed analysis.

The proposed framework demonstrates that mobile ticketing data can be a cost-effective and accessible resource for regional transit agencies seeking to enhance their understanding of passenger behavior without the need for resource-intensive data collection methods. By serving as a decision-support tool, the framework enables agencies to make data-driven decisions to improve service quality, efficiency, and passenger satisfaction. Future research could extend this framework to other transit systems with different characteristics, integrate additional data sources such as real-time vehicle location data, and refine the model to capture real-time changes in ridership patterns. In addition, improving the model to account for sparse data in certain areas or times could further enhance its robustness. The framework also offers a replicable approach that can be adapted by other transit agencies facing similar data constraints, ultimately contributing to more efficient and responsive public transporta-

tion systems.

Authors' Contributions

The authors confirm their contribution to the paper as follows: study conception and design: JO, MA; data collection: MA; algorithm development and coding: MA; analysis and interpretation of results: MA, JO, TO; draft manuscript preparation: MA, JO, TO. All 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.

Acknowledgements

The authors thank Alex Forrest, 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, contributions to the interpretation of results, and feedback significantly enhanced the quality of this study.

References

Mohammed Abdalazeem and Jimi Oke. Extracting Spatiotemporal Bus Passenger Trip Typologies from Noisy Mobile Ticketing Boarding Data. *Data Science for Transportation*, 5, September 2023. doi: 10.1007/s42421-023-00082-x.

Mohammed Abdalazeem and Jimi Oke. Enhanced Seasonal Typology-Informed Transit Trip Chaining via Mobile Boarding and Survey Data. *Data Science for Transportation*, 6(3): 19, September 2024. ISSN 2948-1368. doi: 10.1007/s42421-024-00108-y.

Azalden A. Alsger, Mahmoud Mesbah, Luis Ferreira, and Hamid Safi. Use of Smart Card Fare Data to Estimate Public Transport Origin–Destination Matrix. *Transportation Research Record*, 2535(1):88–96, January 2015. ISSN 0361-1981. doi: 10.3141/2535-10.

Moshe E Ben-Akiva, Peter P Macke, and Poh Ser Hsu. Alternative Methods to Estimate Route-Level Trip Tables and Expand On-Board Surveys. *Transportation Research Record*, page 11, 1985.

Z. Cheng, M. Trepanier, and L. Sun. Inferring trip destinations in transit smart card data using a probabilistic topic model. *Bibliothèque et Archives Canada*, October 2019.

Alex Cui. *Bus Passenger Origin-Destination Matrix Estimation Using Automated Data Collection Systems*. Thesis, Massachusetts Institute of Technology, 2006.

Wei Fan and Zhen Chen. Estimation of Origin-Destination Matrix and Identification of User Activities Using Public Transit Smart Card Data. *Center for Advanced Multimodal Mobility Solutions and Education*, August 2018.

Jason B. Gordon. *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, 2012.

Di Huang, Jun Yu, Shiyu Shen, Zhekang Li, Luyun Zhao, and Cheng Gong. A Method for Bus OD Matrix Estimation Using Multisource Data. *Journal of Advanced Transportation*, 2020:1–13, March 2020. ISSN 0197-6729, 2042-3195. doi: 10.1155/2020/5740521.

Etikaf Hussain, Ashish Bhaskar, and Edward Chung. Transit OD matrix estimation using smartcard data: Recent developments and future research challenges. *Transportation Research Part C Emerging Technologies*, 125, April 2021. doi: 10.1016/j.trc.2021.103044.

Abdalazeem A Mohammed and Jimi Oke. Origin-destination inference in public transportation systems: A comprehensive review. *International Journal of Transportation Science and Technology*, 12(1):315–328, March 2023. ISSN 2046-0430. doi: 10.1016/j.ijtst.2022.03.002.

Neema Nassir, Alireza Khani, Sang Gu Lee, Hyunsoo Noh, and Mark Hickman. Transit Stop-Level Origin–Destination Estimation through Use of Transit Schedule and Automated Data Collection System. *Transportation Research Record*, 2263(1):140–150, January 2011. ISSN 0361-1981. doi: 10.3141/2263-16.

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

Pierre Robillard. Estimating the O-D matrix from observed link volumes. *Transportation Research*, 9(2):123–128, July 1975. ISSN 0041-1647. doi: 10.1016/0041-1647(75)90049-0.

Stan Salvador and Philip Chan. FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space. page 11, 2004.

Gabriel E. Sánchez-Martínez. Inference of Public Transportation Trip Destinations by Using Fare Transaction and Vehicle Location Data: Dynamic Programming Approach. *Transportation Research Record*, 2652(1):1–7, January 2017. ISSN 0361-1981. doi: 10.3141/2652-01.

Yanshuo Sun and Ruihua Xu. Rail Transit Travel Time Reliability and Estimation of Passenger Route Choice Behavior: Analysis Using Automatic Fare Collection Data. *Transportation Research Record*, 2275(1):58–67, January 2012. ISSN 0361-1981. doi: 10.3141/2275-07.

A. G. Wilson. The Use of the Concept of Entropy in System Modelling. *Operational Research Quarterly (1970-1977)*, 21(2):247–265, 1970. ISSN 0030-3623. doi: 10.2307/3008157.

Khatun E Zannat and Charisma F. Choudhury. Emerging Big Data Sources for Public Transport Planning: A Systematic Review on Current State of Art and Future Research Directions. *Journal of the Indian Institute of Science*, 99(4):601–619, December 2019. ISSN 0019-4964. doi: 10.1007/s41745-019-00125-9.