Transitway Impacts Research Program
Report #25 in the series

Transfer Behavior and Off-Peak Commutes

Alireza Khani, Kwangho Baek
October 2024

CTS Report 24-06

CENTER FOR TRANSPORTATION STUDIES
UNIVERSITY OF MINNESOTA

Technical Report Documentation Page

1. Report No. CTS 24-06 / TIRP #25	2.	3. Recipients Accession No.
4. Title and Subtitle Transfer Behavior and Off-Peak Commutes		5. Report Date October 2024
		6.
7. Author(s) Kwangho Baek, Alireza Khani		8. Performing Organization Report No.
9. Performing Organization Name and Address Civil, Environmental, and Geo- Engineering University of Minnesota 500 Pillsbury Drive SE Minneapolis, MN 55455		10. Project/Task/Work Unit No. CTS #2022011
		11. Contract (C) or Grant (G) No.
12. Sponsoring Organization Name and Address Center for Transportation Studies University of Minnesota 440 University Office Plaza 2221 University Avenue SE Minneapolis, MN 55414		13. Type of Report and Period Covered Final Report
		14. Sponsoring Agency Code

15. Supplementary Notes

https://www.cts.umn.edu/publications/researchreports/

https://www.cts.umn.edu/programs/tirp

16. Abstract (Limit: 250 words)

To improve transit service for off-peak travelers, an essential yet often underrepresented group, and promote social equity, this study examines off-peak transit commutes and transfers, with a focus on the transitway system in the Twin Cities. The research contrasts off-peak and peak travel behaviors using an onboard survey (OBS) from 2016 and automatic fare collection (AFC) data from 2018 to 2023. The initial analysis involved clustering trips from OBS into 16 regional zones and creating origin-destination matrices to explore spatial travel patterns. Key findings include longer peak-time trips (8.51 miles) compared to off-peak trips (5.74 miles) and a higher concentration of non-work trips during off-peak times. The study also reveals that off-peak trips are more dispersed geographically. In the second phase, path choice sets were generated for each respondent from OBS, and logistic regression models were used to analyze preferences for transitway versus bus-only routes. The results indicated a strong preference for transitways, with 60% of passengers opting for them over buses when travel times were equal. Finally, AFC data was integrated with OBS using machine learning techniques to examine long-term trends, including the impact of the COVID-19 pandemic. Post-pandemic data show an increase in off-peak commutes and transit trips with transfers despite an overall decline in transfers. This study provides insights into evolving transit usage behaviors and highlights the importance of the transitway system in facilitating efficient travel.

17. Document Analysis/Descriptors Surveys, Smart cards, Off peak periods, Transfers, Bus rapid transit, Light rail transit, Machine learning, Cluster analysis		18. Availability Statement No restrictions. Document available from: National Technical Information Services, Alexandria, Virginia 22312	
19. Security Class (this report) Unclassified	20. Security Class (this page) Unclassified	21. No. of Pages 83	22. Price n/a

TRANSFER BEHAVIOR AND OFF-PEAK COMMUTES

FINAL REPORT

Prepared by:

Kwangho Baek
Alireza Khani
Department of Civil, Environmental, and Geo- Engineering
University of Minnesota

OCTOBER 2024

Published by:

Center for Transportation Studies
University of Minnesota
440 University Office Plaza
2221 University Avenue SE
Minneapolis, MN 55414

ACKNOWLEDGMENTS

The authors of this research thank the members of the Technical Advisory Panel (TAP) who provided data, insightful comments, and agency-specific thoughts on the intermediate and final products of this research project, including Emma Lucken (Ramsey County), Jonathan Ehrlich & Bradley Bobbitt (Metropolitan Council), John Levin (Metro Transit), Sara Allen (Washington County), and Cathy Gold (Hennepin County).

TABLE OF CONTENTS

CHAPTER 1: Introduction 1

CHAPTER 2: Off-Peak Trips and Travelers 4

2.1 Survey Data and Zoning 4

2.1.1 Fundamental Trip Characteristics Revealed from the Survey Data 4

2.1.2 Zoning the Twin Cities Metropolitan Area 8

2.2 Off-Peak Travel Demographics Compared to Peak Trips 14

2.2.1 Mosaic Plots 14

2.2.2 Colored Origin-Destination (OD) Matrices 18

CHAPTER 3: The Transitway’s Impact on Path Choices 23

3.1 Identifying and Pairing Paths for Each Survey Respondent 23

3.1.1 Applying Schedule-Based Transit Shortest Pathfinding Algorithm 24

3.1.2 Pairing the Identified Paths to Create the Path Choice Dataset 28

3.2 Logistic Regression Analysis 30

CHAPTER 4: Off-Peak Transit Demand Analysis with AFC 35

4.1 Trip Chaining 37

4.2 Trip Purpose Inference Using Machine Learning 40

4.2.1 Training and Selecting Machine Learning Algorithms 41

4.2.2 Commute Trip Analysis from the Neural Network Model’s Inference Result 42

CHAPTER 5: Conclusion 50

References 52

Appendix A

LIST OF FIGURES

Figure 1.1 The Twin Cities transit network and transitways (2023) 2

Figure 2.1 Origin/destination place types by time of departure 5

Figure 2.2 Travel distance distributions by time of departure 6

Figure 2.3 Trip counts by route types and time of departure 6

Figure 2.4 HDBSCAN results (AM peak) 7

Figure 2.5 HDBSCAN results (non-AM-peak) 8

Figure 2.6 Zoning result from HDBSCAN, a geoprocessing algorithm, and natural/built barriers 10

Figure 2.7 Mosaic plot example and explanation 15

Figure 2.8 An illustration of the chi-square test for independence for two variables 15

Figure 2.9 Mosaic plots: transitway inclusion in the surveyed path 17

Figure 2.10 Mosaic plots: income group 18

Figure 2.11 Matrix plots: baseline 19

Figure 2.12 Matrix plots: Off-peak proportions of trips with and without transitway routes 20

Figure 2.13 Matrix plots: trip counts by times of departure for peak and off-peak times 20

Figure 2.14 Matrix plots: using trips started from the respondents' home 21

Figure 2.15 Matrix plots: using trips destined to the respondents' work 21

Figure 2.16 Matrix plots: using trips that are not destined to the respondents' home or work 22

Figure 3.1 An example of a path choice set a survey respondent might have had 23

Figure 3.2 A transit path's actual travel stages and path cost 24

Figure 3.3 An illustration of the enhanced SBSP (Tomhave et al., 2022) 25

Figure 3.4 An example of the enhanced SBSP execution for a sample: two transit paths identified 27

Figure 3.5 Average time attributes of the path choice dataset by their recorded path choice 28

Figure 3.6 Passenger grouping depending on their paths' transfer inclusion 29

Figure 3.7 Path choice proportion of the paired path choice dataset by group 30

Figure 3.8 Logistic regression curves for the total travel time and walking time differences.................... 32

Figure 3.9 Disaggregated logistic regression curves by travel context and demographic.......................... 33

Figure 4.1 Raw AFC data count by day... 37

Figure 4.2 What trip chaining algorithms can infer 38

Figure 4.3 Data retrieval performance tracked for applying the trip chain algorithm 39

Figure 4.4 Transfer trip proportion and the share of transitways in transfer trips 40

Figure 4.5 Illustration of the ML application: trip purpose augmentation to the trip chain results 41

Figure 4.6 Transfer trip proportions by destination location and time of day across periods 43

Figure 4.7 The proportion of inferred commute trips from total by workplace location and time of day 44

Figure 4.8 The daily average inferred counts of commute trips from the neural network model............ 45

Figure 4.9 Commute trip proportion destined to each grid by time 46

Figure 4.10 Time of day concentration among commute trips destined to each grid 47

Figure 4.11 A comparison of peak- and off-peak dominance of transfers by periods 48

Figure 4.12 A comparison of pre-and post-pandemic transfer count changes by the time of day............ 49

LIST OF TABLES

Table 2.1 Three types of time of departure classifications 4

Table 2.2 Zonal origin/destination trip counts and proportions by time of departure 11

Table 3.1 The applied SBSP algorithm's travel time weights and algorithm thresholds for pathfinding ... 26

Table 3.2 The univariate logistic regression models result 31

Table 4.1 Datasets' count and representativeness comparison: OBS 2016 and AFC 35

Table 4.2 The validity evaluation of the trained machine learning models 42

Table 4.3 Commute proportion and AM-peak commute proportion by zone 47

EXECUTIVE SUMMARY

This research employed an on-board survey (conducted in 2016) and automatic fare collection (AFC) data (2018–2023) to scrutinize off-peak transit commutes and transfers, particularly emphasizing the transitway system. Initially, the study focused on analyzing the survey, conducting a descriptive analysis that compared Twin Cities transit users' off-peak travel behaviors with peak-time behaviors. Trips were clustered into 16 zones in the region, and mosaic plots were employed to inspect the relationships between specific demographic variables and the time of departure by zone. Additionally, detailed origin-destination trip matrices were presented to illustrate spatial travel behavior patterns. The key findings from this initial analysis are summarized below:

- Peak-time transit travels exhibit a longer average distance (8.51 mi) than off-peak transit travels (5.74 mi), scrutinized by a t-test.
- Home-based (92%) and commute (72%) trips dominate AM-peak, while trip purposes like shopping and social visits are more inclined toward off-peak times (62% and 58%, respectively).
- Clustering results highlight that off-peak trip locations are more scattered than those of peak.
- Matrix plots indicate that transitway trips and non-home or non-work trip purposes are concentrated more during off-peak hours.

The subsequent analysis recovered transitway-including and bus-only transit path options for each eligible respondent, pairing the paths to construct a path choice set. Choice results were then analyzed, with logistic regression models mapping travel time differences to transitway path choice probabilities. The results validated passengers' preferences for transitways over bus-only options. Furthermore, the analysis highlighted variations in this preference across different groups. The comparison results are summarized below:

- Recovered path choices suggest that transitways facilitate transfers, as 23% of people took a transfer-including transitway path even when they had a choice to take a direct bus path.
- Logistic regression findings indicate that 60% of passengers would choose transit paths with transitway routes over bus-only alternatives when travel times were equal.
- Commuters, especially peak-time transit users, displayed less sensitivity to modes; they tended to prioritize travel time savings over the inclusion of a transitway route in their journey.

Finally, the research team enhanced and integrated AFC data with the survey using machine learning techniques. This facilitated the analysis of long-term commute behaviors from 2018 to 2023, capturing periods without and with the impact of the COVID-19 pandemic. The inferred commute trips from the selected model were visualized through map representations. Key findings from the inference include:

- About a third of transfers were associated with transitways.
- A notable increase in the proportion of off-peak commutes was observed after the pandemic: the proportion of trips heading to the downtowns of Minneapolis and St. Paul increased from 28% to 40% and to other areas from 15% to 23%.

- Before the pandemic, 68% of transit commutes to workplaces occurred during the AM peak. This number decreased to 46% post-pandemic.
- After the pandemic, the total number of transfers decreased, but the proportion of transit trips with transfers increased.

CHAPTER 1: INTRODUCTION

Transit services are indispensable in urban areas across the United States, catering to a diverse population with varying travel requirements. Despite the expansive geographical coverage and extended operational hours of transit services, various constraints often lead to indirect routes, necessitating transfers between origins and destinations. In the Twin Cities area, the launch of efficient transitway services (see **Figure 1.1**), consisting of light rail transit (LRT) and bus rapid transit (BRT), is believed to have enhanced the transit experience by promoting faster, more frequent, and reliable services. With consistent headways throughout the day, these transitways play a crucial role in accommodating off-peak trips and the associated transfers, mainly when express bus services are limited or non-existent.

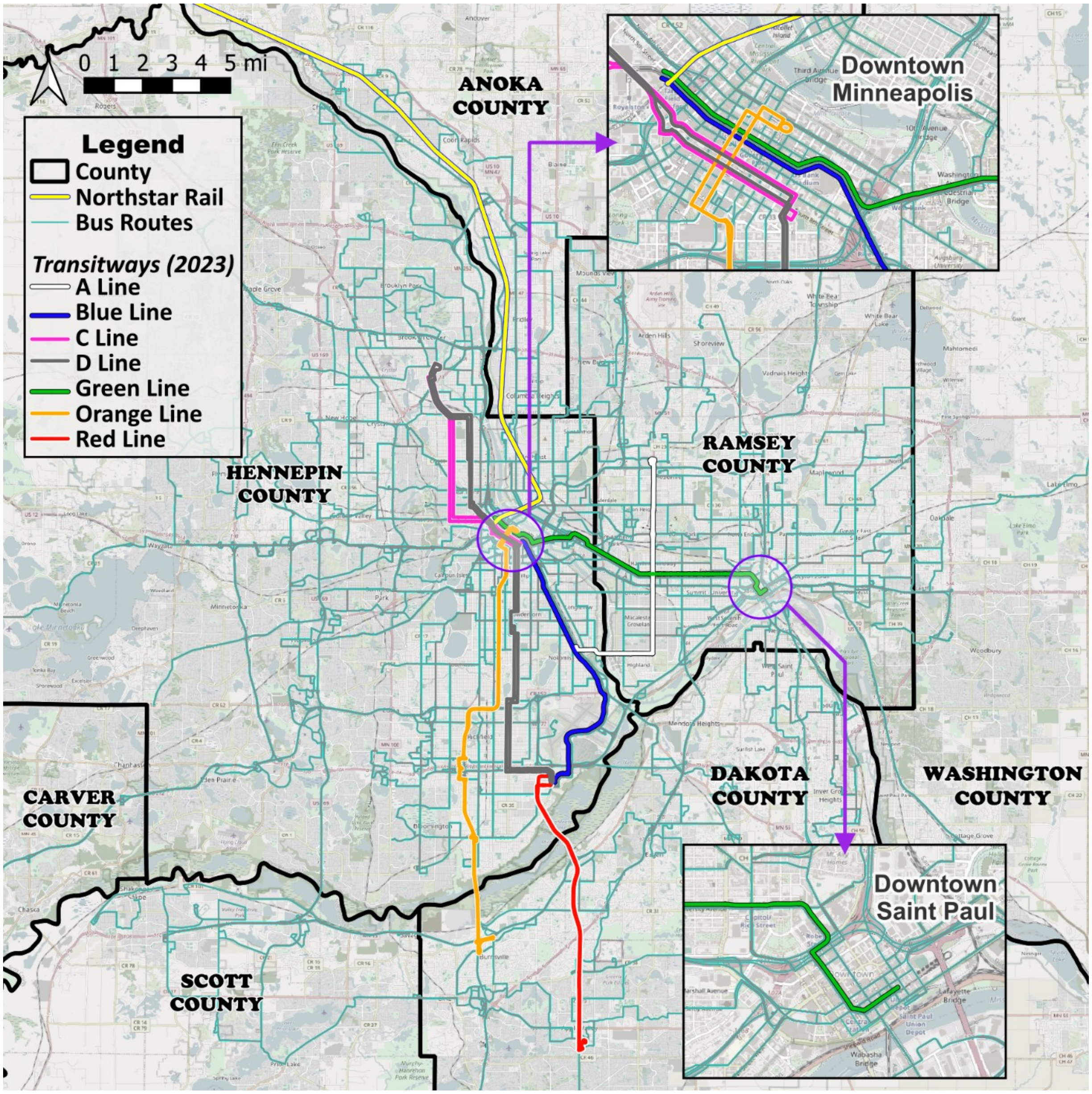


Figure 1.1 The Twin Cities transit network and transitways (2023)

Transit investments have traditionally prioritized the needs of peak-time commuters, with operational decision variables such as route itineraries, fleet size, and staffing tailored to accommodate this demographic (Ceder, 2016; Vuchic, 2005). However, to develop a comprehensive understanding of underrepresented travel patterns, it is crucial to scrutinize the behavior of riders during off-peak hours and examine how transitways support these journeys. This analysis becomes particularly efficient considering a meta-analysis by Hensher (2008) highlighting a higher demand elasticity for off-peak transit trips. This suggests that directing investments toward enhancing off-peak transit utility may yield more effective results, especially in cities with established transit networks like the Twin Cities.

Furthermore, studies suggest that off-peak transit trips are often linked to the commutes of non-white-collar, essential workers and transit-dependent populations, and the COVID-19 pandemic has highlighted the need for a closer examination of these trips (Parker et al., 2021). On the other hand, other studies found that people who were peak-time travelers before the pandemic, mostly choice riders, changed their commute patterns during the pandemic (Wang et al., 2021).

In addition to off-peak travel analysis, this project delves into the other expected impact of transitways in facilitating transfers, exploring passenger perceptions evidenced by revealed path choices. It has been proven that such an approach is helpful for capturing and quantifying transit users' preferences toward specific features of transit. For example, Guo & Wilson (2011) generated path choice sets from surveyed datasets to assess the transfer cost of the London Underground. Raveau et al. (2014) used London's and Santiago's survey data for route choice model estimation to examine the impact of various elements on passengers' metro transfer. Their results revealed that participants were more sensitive to generalized travel cost (the sum of perceived travel time and inconvenience) than absolute travel time, and improving comfort at the transfer point will increase transit ridership. However, most of these studies have focused on user perceptions of different modes (e.g., light rail versus bus routes), with limited research addressing historical path choices.

More recently, automatic fare collection (AFC) data has become an excellent source for analyzing transit users' transfer behavior. Zhao et al. (2017) proposed a method using AFC that identifies major transfer points that need to be prioritized for facility management. They estimated each time component—walking time, delay, and waiting time—of transfers from the metro to the bus. Zhu et al. (2020) and Hörcher et al. (2017) developed passenger-to-train-assignment algorithms to recover platform and in-vehicle conditions, including transfer station crowding. Webb et al. (2020) developed a trip-chaining algorithm to infer each passenger's trajectory and transfer stages from AFC data without alighting records. However, studies that used AFC data to investigate trips with specific purposes (e.g., commute) are limited due to the absence of information in the data.

This project employs onboard transit rider survey and AFC data for off-peak commute and transfer analyses. First, descriptive statistics are presented, and statistical tests are employed to investigate peak- and off-peak heterogeneity by demographics. It is then followed by quantification of riders' perceptions of transitway paths compared to non-transitway paths. Finally, by attaching machine-learning-inferred implicit trip purpose to AFC data, we analyzed long-term commute trips and examined the impact the COVID-19 pandemic had on transit use behaviors in the region.

CHAPTER 2: OFF-PEAK TRIPS AND TRAVELERS

The 2016 transit on-board survey (OBS 2016) data was utilized to analyze off-peak travel behaviors and transfers. The collected dataset features 35,761 transit passenger observations with full-scale travel and user characteristics, including origin, destination, trip purpose, access/egress mode, transfers, and various socio-demographic characteristics. For this and subsequent chapters, we aggregated the data into either two-, three- or five-level time of departure classifications when reporting statistics depending on the relevance/convenience of interpreting the results. The five-level classification is defined with the reference periods suggested in OBS 2016, specifically AM Peak (6 – 9 AM) and PM Peak (3-6:30 PM). A day can be divided into periods before the AM Peak, between the two peaks, and after the PM Peak, so these create a five-level classification. It is then simplified to a three-level classification if we let all three off-peak periods as one class. Finally, by combining the two peaks, we can simply divide a day into peak and off-peak. The three types of time of departure classification are summarized in **Table 2.1**.

Table 2.1 Three types of time of departure classifications

Time of Departure	**04:00 – 06:00**	**06:00 – 09:00**	**09:00 – 15:00**	**15:00 – 18:30**	**18:30 – 04:00**
Five-level classification	Early Morning	AM Peak	Midday	PM Peak	Night
Three-level classification	Off-peak	AM Peak	Off-peak	PM Peak	Off-peak
Two-level classification	Off-peak	Peak	Off-peak	Peak	Off-peak

2.1 SURVEY DATA AND ZONING

This section outlines key features of the surveyed trips, leveraging the precise origin/destination locations recorded in each survey response. We used this spatial information to conduct geostatistical cluster analysis and zoning for more in-depth examinations.

2.1.1 Fundamental Trip Characteristics Revealed from the Survey Data

In transportation economics, the purpose of a trip plays a pivotal role in individuals' decision-making regarding whether to embark on a specific journey. Specifically, a trip happens if and only if a person's anticipated benefits derived from the trip outweigh the associated travel costs, encompassing both time and monetary expenditures (de Dios Ortúzar & Willumsen, 2011). Transit options vary across different times of the day, and the fluctuating traffic congestion further contributes to the dynamic nature of travel costs. Consequently, a fundamental aspect of analyzing a region's transit travel characteristics involves understanding users' trip purposes at different times.

Figure 2.1 provides a visual representation of colored tables depicting origin/destination place types defined in OBS 2016 across five times of departure. The surveyed counts, normalized to rides per hour, emphasize that trips during the AM peak from home and PM peak to home exhibit the highest values in each table. Subsequently, PM peak trips from work and AM peak trips to work follow those. In off-peak periods, home-originated and work-destined midday trips stand out as the two most populous instances, indicating a significant number of off-peak trips facilitated by the transit system. Personal trips such as shopping, social visits, and recreational outings appear relatively evenly distributed, with a higher prevalence in later times (midday, PM Peak, and night). Similarly, student trips to university or college are spread across different times of the day, but with a concentration at earlier times (AM peak and midday).

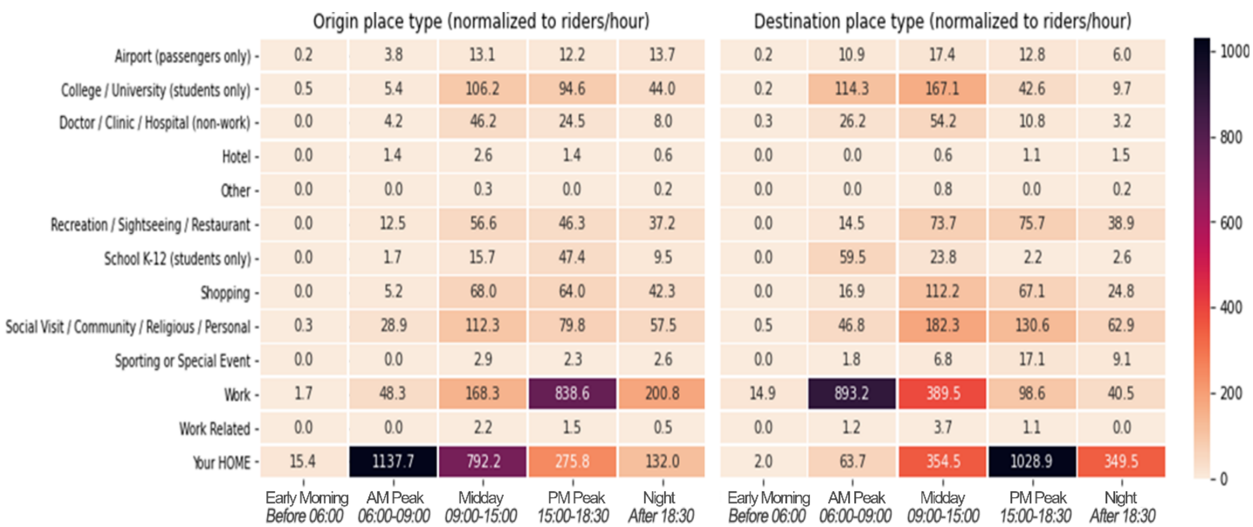


Figure 2.1 Origin/destination place types by time of departure

Figure 2.2 depicts the distribution of travel distances (Euclidean distance from origin to destination) for peak-time and off-peak-time travelers, with the 90th and 95th quantiles. Histograms reveal a trend where the average travel distance for off-peak trips is shorter than that for peak-time trips. A two-sample T-test confirms a statistically significant difference in mean travel distances between peak and off-peak time travelers. This discrepancy can be again attributed to transportation economics, wherein the anticipated benefits of a commute trip, concentrated at peak times, outweigh the high travel cost or distance. Alternatively, the difference could have been due to the higher availability of express buses during peak times, covering extensive distances, contributing to decreased travel costs.

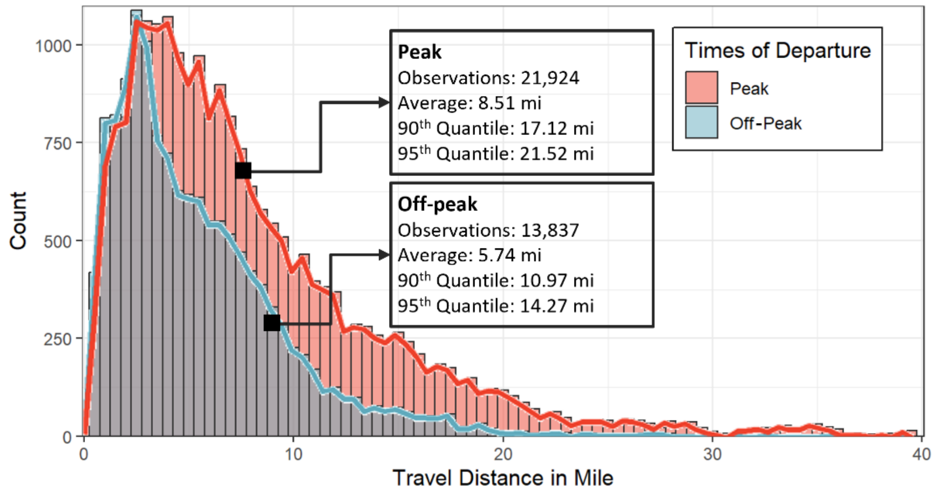


Figure 2.2 Travel distance distributions by time of departure

Survey responses further unveil a noticeable incongruity between times of departure and route types, as depicted in **Figure 2.3**. Express buses exhibit a remarkable concentration during peak times, while other route types do not display such pronounced patterns. Notably, at the time of the survey, there were only two BRT routes in the Twin Cities. One of these routes, the Metro A Line, commenced its service during the survey period, resulting in few observations for BRT.

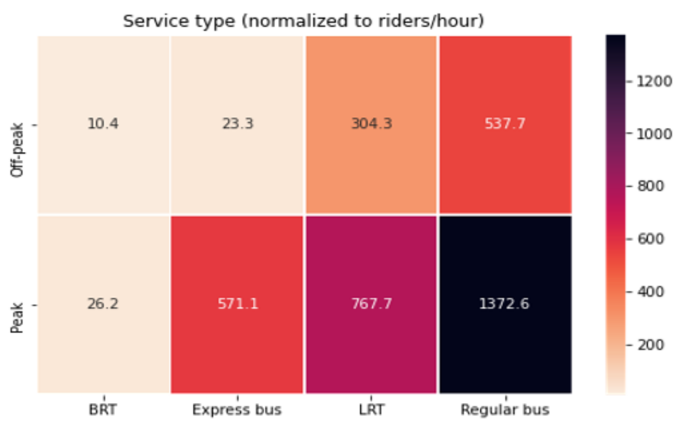


Figure 2.3 Trip counts by route types and time of departure

The spatial aspect of each trip, involving the origin/destination points, allows for identifying popular geostatistical clusters. Employing the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN; Campello et al., 2015) algorithm, **Figure 2.4** and **Figure 2.5** present the clustering results for AM Peak and Non-AM Peak times, respectively.

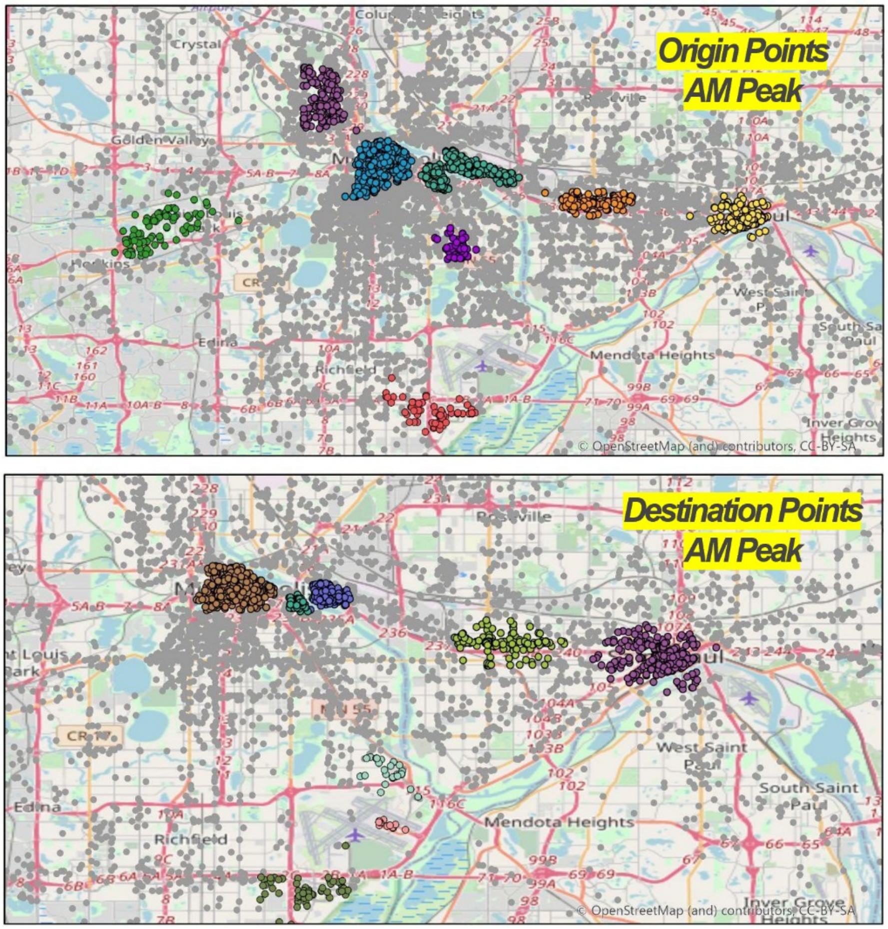


Figure 2.4 HDBSCAN results (AM peak)

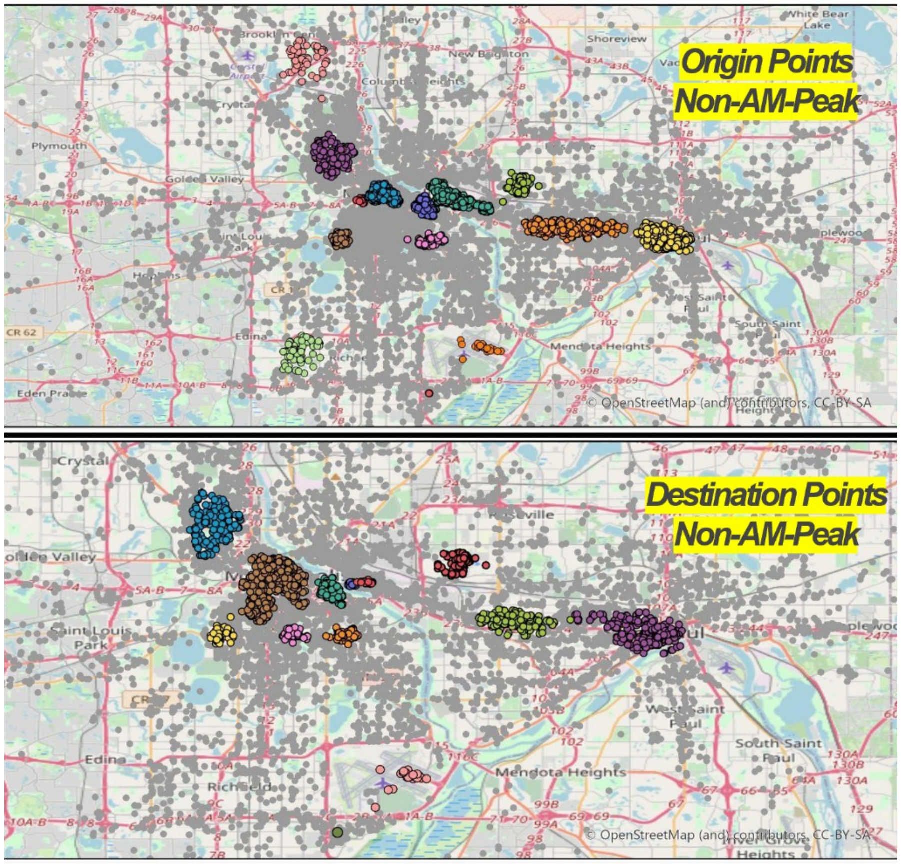


Figure 2.5 HDBSCAN results (non-AM-peak)

The maps reveal a concentration of destination points in downtown areas during AM peak trips, contrasting with more scattered origin clusters. Non-AM peak trips exhibit a less concentrated pattern. Notably, more clusters are identified in areas adjacent to transitway lines, indicating their influence on trip patterns.

2.1.2 Zoning the Twin Cities Metropolitan Area

To gain a comprehensive understanding of transit riders at a smaller geographical level, it is customary to map individual points to named areas or zones. While administrative borders, such as census tracts or county subdivisions, are often employed in zoning, they may not adequately capture transit services' linear, fixed-routed nature. Therefore, a novel zoning approach that effectively delineates in-zone homogeneity and between-zone heterogeneity in transit usage is essential for spatial analyses of transit travel behavior. The geoprocessing tool "Build Balanced Zones" in ArcGIS Pro employs a genetic algorithm to create spatially contiguous zones based on input points and user-defined criteria (ESRI, 2021). Utilizing this tool with the survey's origin and destination points combined, we initially generated 15 zones. Subsequent manual adjustments considered HDBSCAN results from the previous subsection and accounted for built/natural barriers such as railroads/highways, existing transit routes, and rivers. The outcome of this zoning process, featuring 15 zones, is illustrated in **Figure 2.6**. Notably, the added last zone (Zone 16: *PNR Dominant Points*) represents suburban regions where regular public transit services are unavailable within 1.1 miles. Zone details, categorized by the time of departure counts and proportions, are summarized in **Table 2.2**.

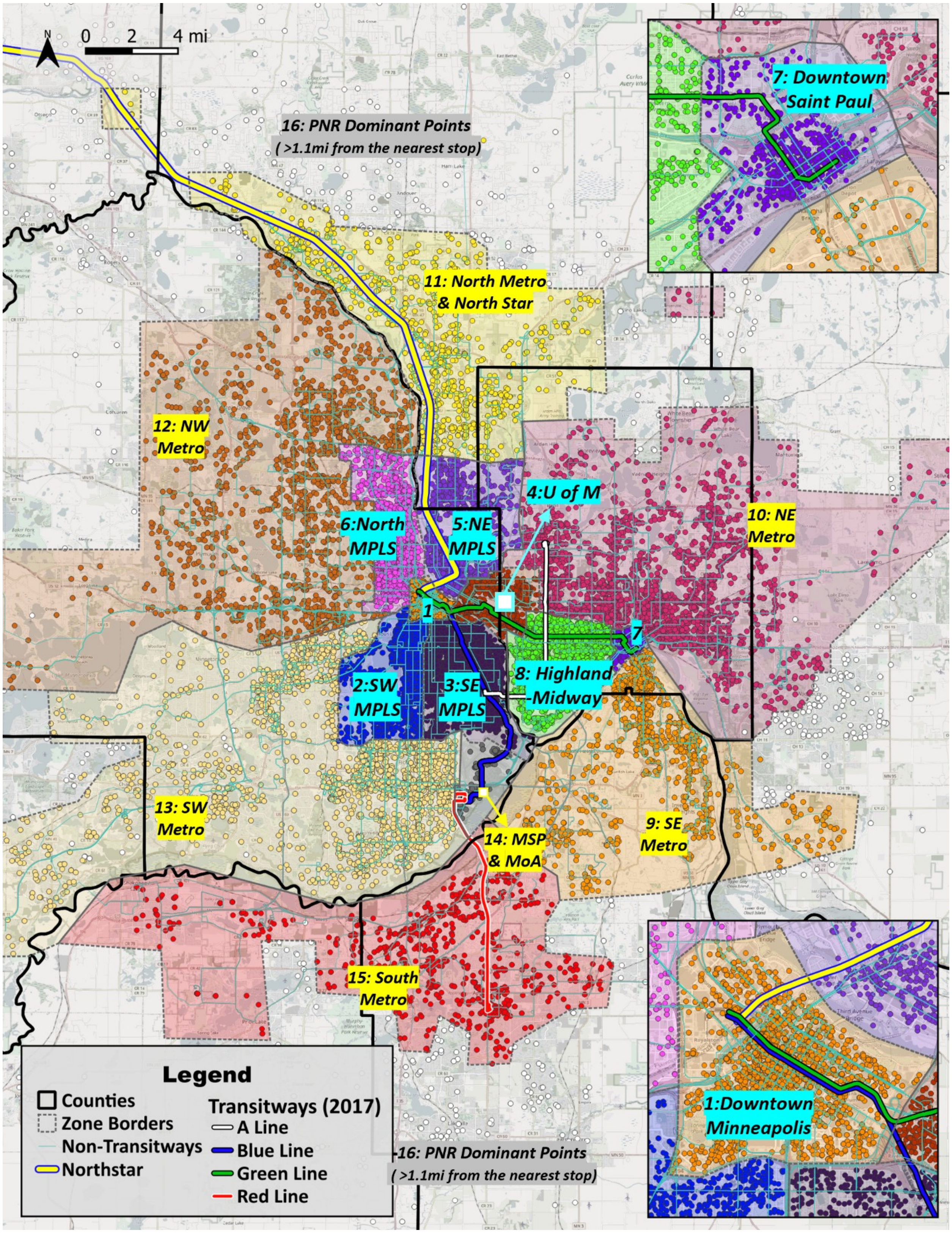


Figure 2.6 Zoning result from HDBSCAN, a geoprocessing algorithm, and natural/built barriers

Table 2.2 Zonal origin/destination trip counts and proportions by time of departure

Zone	Number of Origin Points in the Zone				Number of Destination Points in the Zone				Total (100%)
	AM Peak	Off-Peak	PM Peak	**Subtotal**	AM Peak	Off-Peak	PM Peak	**Subtotal**	
01: Downtown Minneapolis (DTMP)	844 (6.6%)	1,680 (13.2%)	1,566 (12.3%)	**4,090 (32.2%)**	6,192 (48.8%)	2,103 (16.5%)	332 (2.6%)	**8,627 (67.8%)**	**12,717**
02: Southwest Minneapolis (SWMP)	1,252 (26.1%)	1,038 (21.6%)	300 (6.2%)	**2,590 (54.0%)**	614 (12.8%)	1,046 (21.8%)	548 (11.4%)	**2,208 (46.0%)**	**4,798**
03: Southeast Minneapolis (SEMP)	1,648 (25.4%)	1,552 (23.5%)	432 (6.7%)	**3,602 (55.5%)**	1,100 (17.0%)	1,346 (20.7%)	440 (6.8%)	**2,886 (44.5%)**	**6,488**
04: University of Minnesota (UMN)	1,116 (14.1%)	1,938 (24.5%)	524 (6.6%)	**3,578 (45.2%)**	2,178 (27.5%)	1,971 (24.9%)	184 (2.3%)	**4,333 (54.8%)**	**7,911**
05: Northeast Minneapolis (NEMP)	802 (29.6%)	581 (21.4%)	152 (5.6%)	**1,535 (56.7%)**	354 (13.1%)	566 (20.9%)	254 (9.4%)	**1,174 (43.3%)**	**2,709**

06: North Minneapolis (NMP)	858 (25.0%)	830 (24.2%)	182 (5.3%)	**1,870 (54.5%)**	422 (12.3%)	808 (23.5%)	334 (9.7%)	**1,564 (45.5%)**	**3,434**
07: Downtown Saint Paul (DTSP)	382 (9.8%)	619 (15.9%)	586 (15.1%)	**1,587 (40.8%)**	1,418 (36.4%)	699 (18.0%)	188 (4.8%)	**2,305 (59.2%)**	**3,892**
08: Highland-Midway (HLMD)	1,280 (21.2%)	1,494 (24.8%)	554 (9.2%)	**3,328 (55.2%)**	704 (11.7%)	1,345 (22.3%)	656 (10.9%)	**2,705 (44.8%)**	**6,033**
09: Southeast Metro Area (SEMT)	666 (34.6%)	313 (16.3%)	138 (7.2%)	**1,117 (58.0%)**	288 (15.0%)	283 (14.7%)	238 (12.4%)	**809 (42.0%)**	**1,926**
10: Northeast Metro Area (NEMT)	1,648 (28.9%)	1,120 (19.6%)	446 (7.8%)	**3,214 (56.3%)**	748 (13.1%)	1,027 (18.0%)	716 (12.6%)	**2,491 (44.3%)**	**5,705**
11: North Metro Area & North Star (NMTN)	728 (43.1%)	236 (14.0%)	54 (3.2%)	**1,018 (60.3%)**	192 (11.4%)	234 (13.9%)	244 (14.5%)	**670 (39.7%)**	**1,688**
12: Northwest Metro Area(NWMT)	1,540 (42.4%)	524 (14.4%)	250 (6.9%)	**2,314 (63.7%)**	474 (13.0%)	497 (13.7%)	348 (9.6%)	**1,319 (36.3%)**	**3,633**
13: Southwest Metro Area (SWMT)	1,380	857	288	**2,525**	754	741	594	**2,089**	**4,614**

	(29.9%)	(18.6%)	(6.2%)	**(54.7%)**	(16.3%)	(16.1%)	(12.9%)	**(45.3%)**	
14: (MSP) Airport and Mall of America (AMOA)	220 (7.8%)	724 (25.8%)	172 (6.1%)	**1,116** **(39.8%)**	652 (23.2%)	861 (30.7%)	178 (6.3%)	**1,691** **(60.2%)**	**2,807**
15: South Metro Area (SMT)	634 (47.7%)	274 (20.6%)	28 (2.1%)	**936** **(70.5%)**	120 (9.0%)	224 (16.9%)	48 (3.6%)	**392** **(29.5%)**	**1,328**
16: Park and Ride Dominant Points (PNR)	1,240 (67.4%)	87 (4.7%)	14 (0.8%)	**1,341** **(72.9%)**	28 (1.5%)	86 (4.7%)	384 (20.9%)	**498** **(27.1%)**	**1,839**

■ Peak-time concentration over 1/3 ■ Peak-time concentration under 5% ■ Off-peak concentration over 20% ■ O or D subtotal concentration over 2/3

Generally, suburban zones from Zone 9 to Zone 16 exhibit high concentrations of AM-peak origin trips and a notably low proportion of PM-peak origin trips. In contrast, downtown zones display high concentrations of AM-peak destination trips. For zones adjacent to downtowns (Zones 2–6 and 8) and the *MSP Airport and Mall of America* Zone, off-peak trips constitute more than 20% of each zone's daily total trips. This phenomenon may be attributed to the diverse land use of these zones, encompassing both residential areas and other trip attraction facilities, or to the presence of transitway services offering reliable transportation options for residents. Finally, three zones (Zones 1, 15, 16) exhibit significant travel direction concentration, indicating that either origin trips from these zones or destinations to them constitute more than two-thirds.

2.2 OFF-PEAK TRAVEL DEMOGRAPHICS COMPARED TO PEAK TRIPS

This section employs visualization techniques to compare off-peak and peak trips, focusing on zone-by-zone socio-demographic attributes and travel contexts. They are followed by matrix plots considering both origin and destination zones together, visualized with graduated colored matrices.

2.2.1 Mosaic Plots

Mosaic plots summarize data counts across three categorical variables (Meyer et al., 2006). **Figure 2.7** provides illustrative mosaic plots generated with OBS 2016, where each zone defined in the previous section acts as a block of cells. Then, the block is internally divided by the remaining two variables. The vertical division uniformly represents off-peak on the left and peak time on the right. However, the horizontal division varies among plots, representing different demographic or travel-context-related attributes. Each cell includes the count of the pertaining data aligned vertically so they can be easily compared with each cell's height. With these specifications, each block or zone, whether it is either trip origin or destination, shows if there are peak- and off-peak- discrepancies of the variable being inspected. The title above each block is accompanied by asterisk signs indicating chi-square (χ^2) statistics, with more stars denoting statistically significant discrepancies.

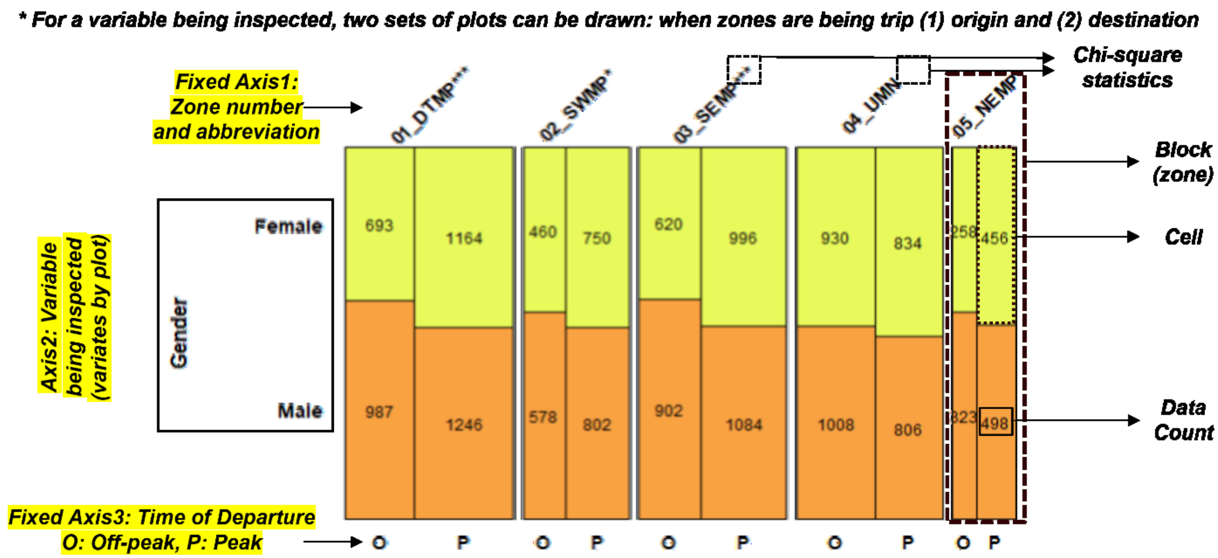


Figure 2.7 Mosaic plot example and explanation

The aforementioned χ^2 statistics can be computed because each block represents a contingency table of two categorical variables. Assume two binary categorical variables, X1 (with possible attribute levels A or B) and X2 (C or D), divide a 100-sample dataset into four cells (see **Figure 2.8**). Marginal proportions are computed by using only one variable's count. For instance, the marginal proportions of A, B, C, and D are 50%, 50%, 60%, and 40%, respectively. Utilizing these marginal proportions, we can estimate the expected count for each cell if they were independent. For example, the expected count for data with attributes X1=A and X2=C is calculated as 50% * 60% * 100 = 30. If there are significant discrepancies between the expected and observed values, the χ^2 value of the table will be large, indicating a relationship or dependence between the two variables, X1 and X2.

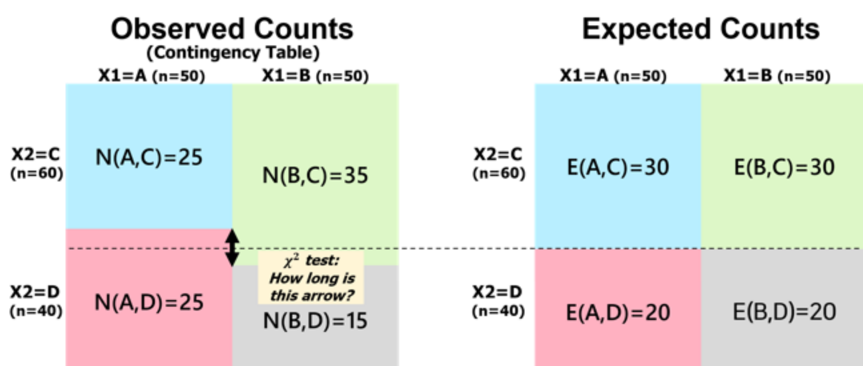


Figure 2.8 An illustration of the chi-square test for independence for two variables

To concisely integrate a substantial volume of information in a mosaic plot, certain multifactorial categorical variables or suitable continuous variables in the survey are restructured into binary or ternary variables. In addition, zone names (with zone numbers) and time of departures (peak: P, off-peak: O) are abbreviated in the plots. Additionally, it is important to emphasize that the number of asterisks following the zone name reflects the statistical significance or P-value of the χ^2 statistic: "*" denotes P<0.05, "**" indicates P<0.01, and "***" signifies P<0.001.

Figure 2.9 shows an example of a pair of mosaic plots where transitway inclusion in the surveyed path is the variable being inspected. The upper panel shows when Downtown Minneapolis (Zone 1), University of Minnesota (4), Downtown St. Paul (7), Highland-Midway (8), Northeast, Northwest, and South Metro Areas (10, 12, 15), and PNR Zone (16) are being the surveyed trip origin, χ^2 statistics captured that peak and off-peak travelers have heterogeneity in the proportion of having transitway in their path. Similarly, we have such heterogeneity for a total of 12 zones when the zones are being trip destinations. Notably, off-peak travelers originating from Downtown Minneapolis exhibit a significantly higher propensity for utilizing transitway services compared to their peak-time counterparts. This distinction may be attributed to the likelihood of peak-time travelers opting for express bus services, connecting the zone to various locations within the metropolitan area. It is noteworthy that the correlation between time of departure and transitway usage is not evident in Southeast Minneapolis (Zone 3) and MSP Airport & Mall of America (Zone 14) when they are both origin and destination places. These are the places where the Metro Blue Line primarily serves.

Likewise, we can draw mosaic plots when the variable being inspected is non-binary. **Figure 2.10** shows mosaic plots where the ternary variable income is the second axis of the plots. The results indicate that in almost every zone, people with different incomes are likely to take their transit services at different times of day. More results and a summary of the mosaic plots can be found in **Appendix A**.

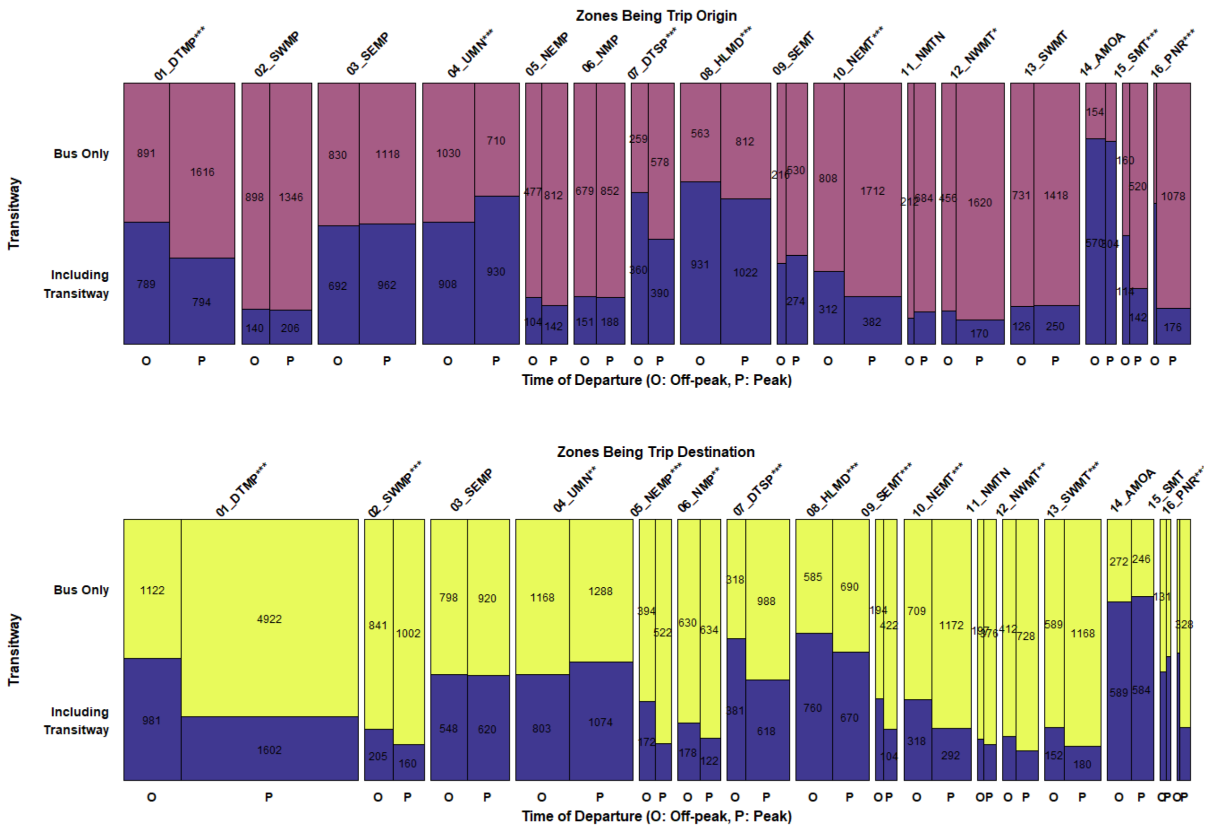


Figure 2.9 Mosaic plots: transitway inclusion in the surveyed path

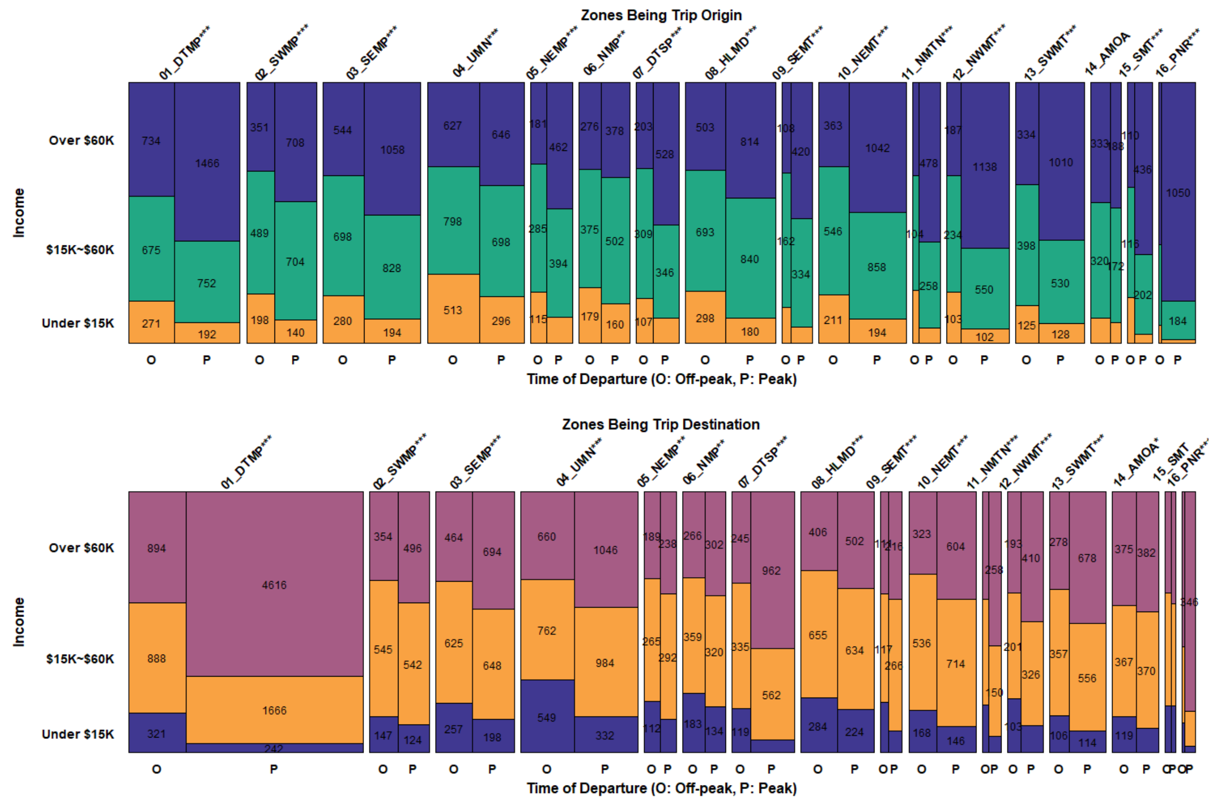


Figure 2.10 Mosaic plots: income group

2.2.2 Colored Origin-Destination (OD) Matrices

The mosaic plots served as aggregated summaries based on either the origin or destination (OD) of the travel. A matrix format is employed for a more comprehensive analysis considering both the OD of transit trips. The matrices, depicted in the following figures, utilize graduated colors, where rows represent trip origin zones and columns represent destination zones. Two matrix types are presented: a trip count matrix and an off-peak trip proportion matrix. The latter is calculated by dividing off-peak trip counts by the total daily trip counts for each OD. The same zoning scheme from the previous section is employed, with cells having peak or off-peak time trip counts of less than five being excluded and colored gray. Additionally, for the trip counts, approximated total daily trip counts (using survey weights) are used to showcase the actual flow of passengers better.

Figure 2.11's left panel presents a matrix plot using the entire factorized dataset. Intrazonal trips along the diagonal of each matrix tend to have maximum values for each row and column. For instance, cells near Zone 1 (Downtown Minneapolis) and Zone 7 (Downtown St. Paul) on the diagonal are typically the darkest, followed by near-diagonal cells representing short-distance trips. The cells in the last row, originating from the PNR zone, and in the last column, destined for the PNR zone, are not very popular OD pairs. However, the first element of each, which is the counts between Downtown Minneapolis (Zone 1) and PNR Zone (16), shows significantly high values. It indicates a tendency for long-distance suburban-to-downtown commuting trips to have more counts than their internal trips. While implications from the off-peak proportion for the aggregated result presented in the right panel are challenging to discern, they serve as a baseline reference for the subsequent results.

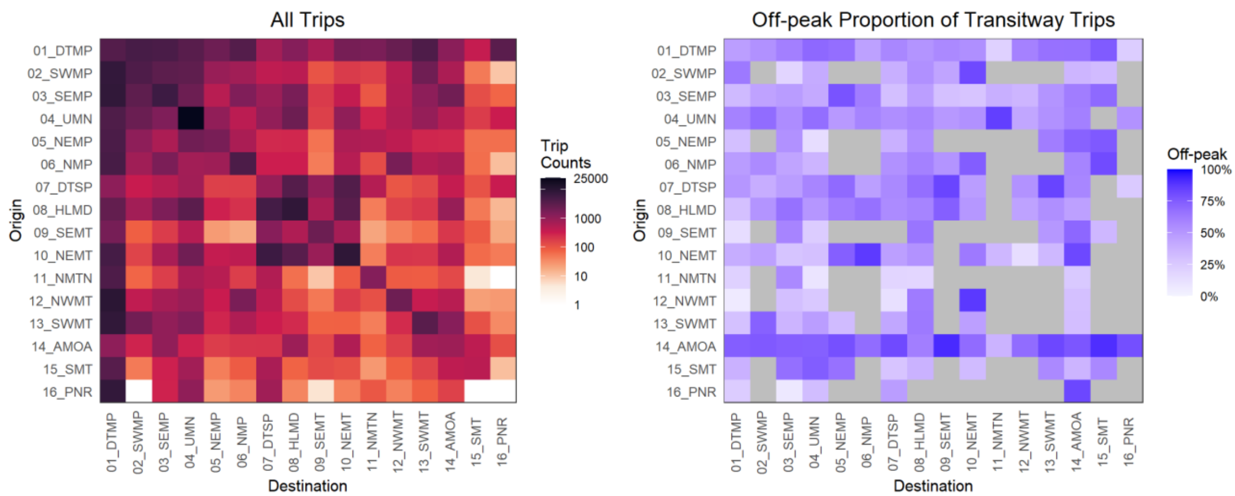


Figure 2.11 Matrix plots: baseline

For example, **Figure 2.12** depicts off-peak proportion matrix plots created separately using the filtered dataset of transitway users and non-transitway users, respectively. Due to limited transitway services linking areas within or between suburban zones (Zones 9–16 except 14), gray cells are more visible in the transitway trip plot. A comparison of the two panels confirms that transitway trips tend to concentrate more on off-peaks, as indicated by relatively darker colors. Noteworthy results include the trips associated with MSP Airport & Mall of America Zone (Zone 14), which exhibits significantly higher off-peak proportions for transitway travel compared to non-transitway travel.

Similarly, **Figure 2.13** illustrates filtered count matrix plots created separately using peak time travelers and off-peak time travelers. We can observe that trips associated with the two downtowns are significantly darker than the other instances for the peak counts, but the differences are less visible for the off-peak matrix.

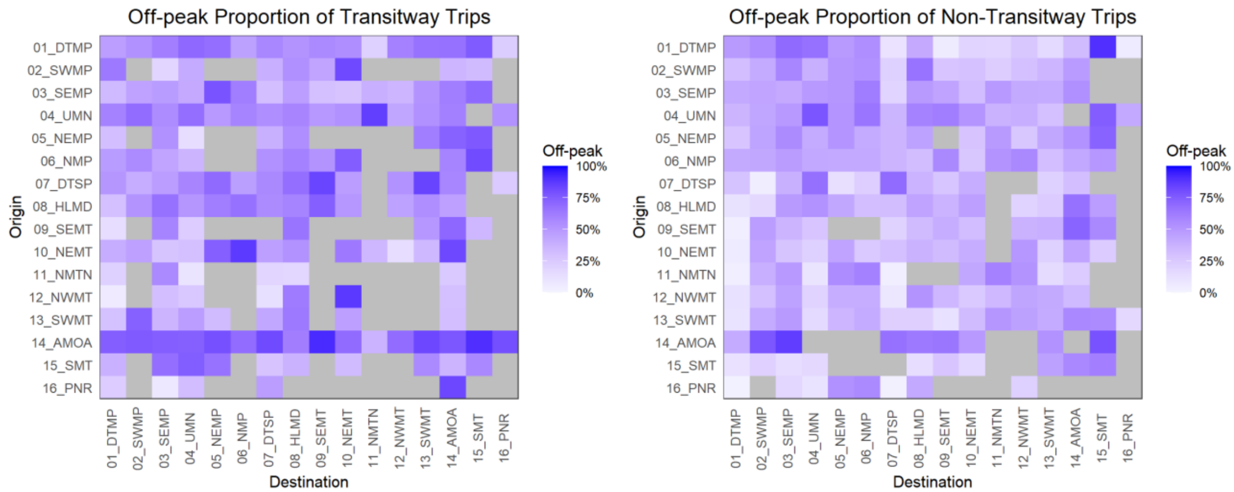


Figure 2.12 Matrix plots: Off-peak proportions of trips with and without transitway routes

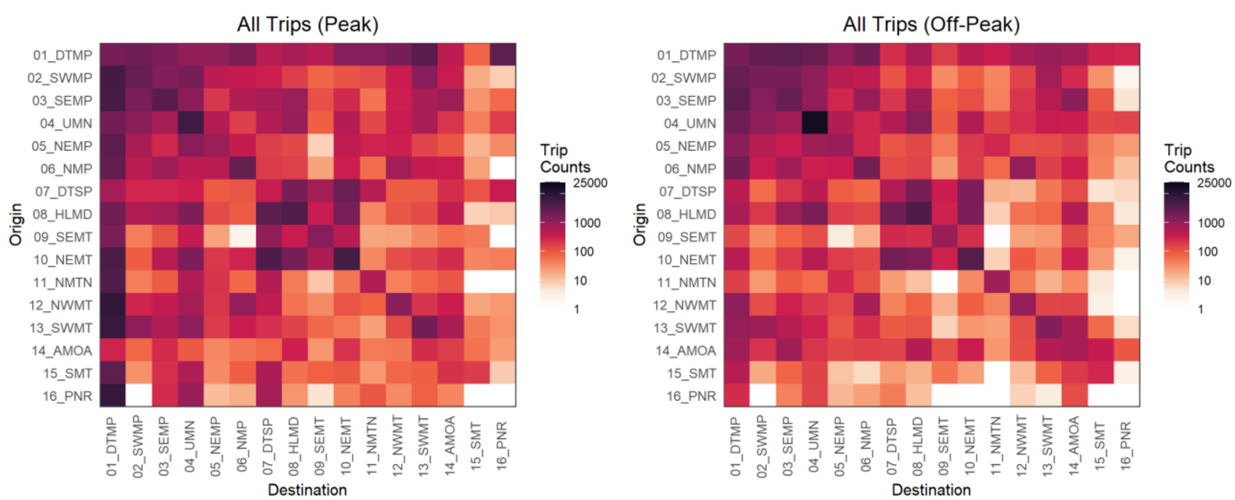


Figure 2.13 Matrix plots: trip counts by times of departure for peak and off-peak times

Finally, we created matrix pairs that only used home-originated trips, to-work commute trips, and trips heading to non-home, non-work destinations (**Figure 2.14** – **Figure 2.16**). The upper two sets of figures reconfirm the commute concentration from suburban homes to work in the downtown area and the low off-peak proportion. The final pair, however, shows relatively evenly distributed spatial count and higher concentrations at off-peak times.

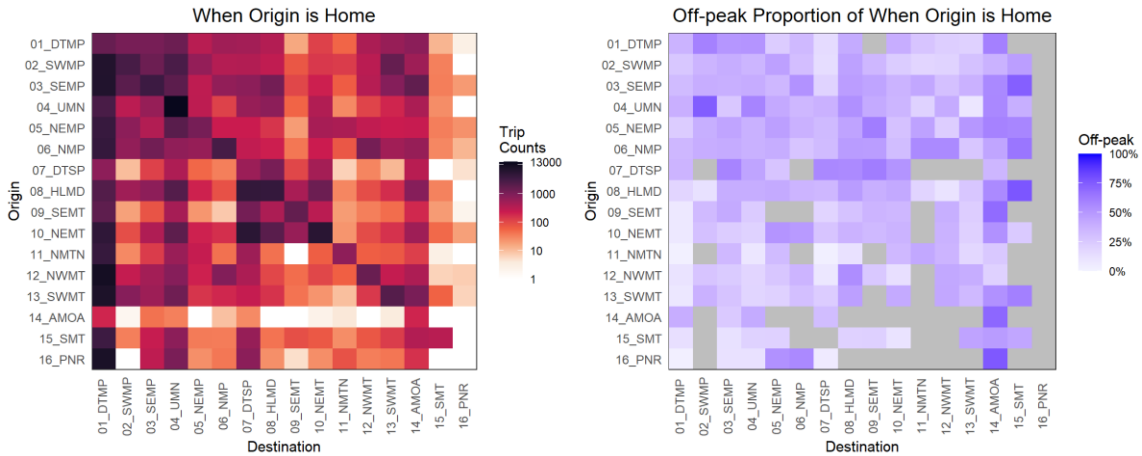


Figure 2.14 Matrix plots: using trips started from the respondents' home

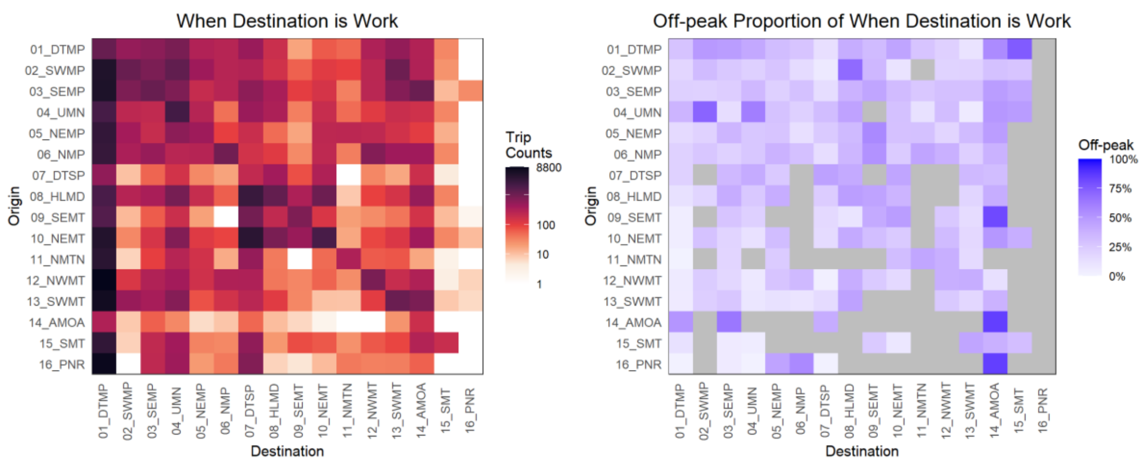


Figure 2.15 Matrix plots: using trips destined to the respondents' work

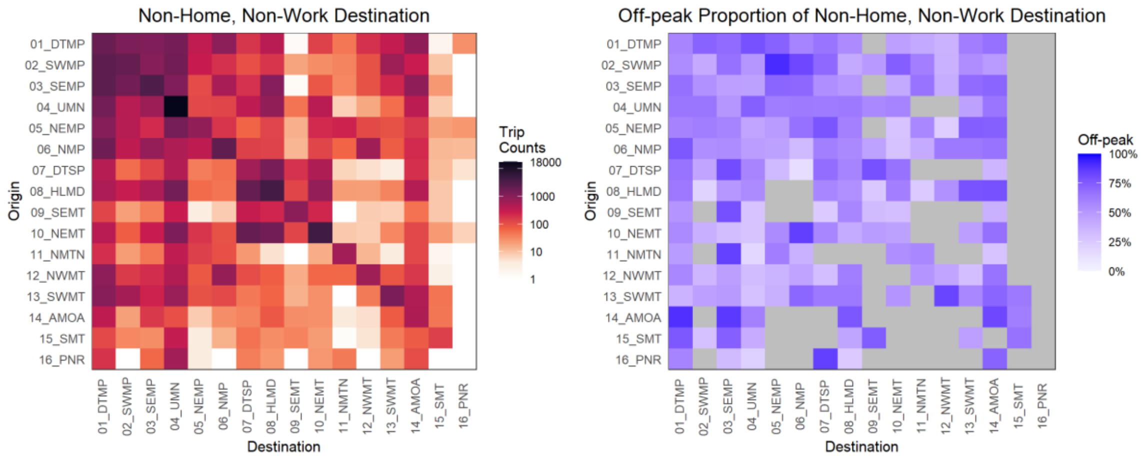


Figure 2.16 Matrix plots: using trips that are not destined to the respondents' home or work

CHAPTER 3: THE TRANSITWAY'S IMPACT ON PATH CHOICES

In this chapter, we aim to investigate the comparative advantages of transitways in transit path choices perceived by general passengers or passengers from different demographic groups. We propose a method that generates, pairs, and compares plausible transit path options. The objectives of the above are to identify whether including transitway route(s) in people's transit paths would compensate for other disutilities (e.g., increased travel time) and, if so, by how much for different groups.

3.1 IDENTIFYING AND PAIRING PATHS FOR EACH SURVEY RESPONDENT

Suppose a record of OBS 2016 suggests the passenger took a transit path on Thursday at 3 PM comprising Metro Blue Line and Route 14, traveling from Downtown Minneapolis to Lake Hiawatha. We can imagine that the passenger may have had alternative options. For instance, given the same origin, destination, and time of departure input, let's assume that the transit trip planner (e.g., Google Maps) recommended two paths, as illustrated in **Figure 3.1**, where the latter option is a direct bus path using Route 22. The two paths collectively constitute the passenger's transit path choice set.

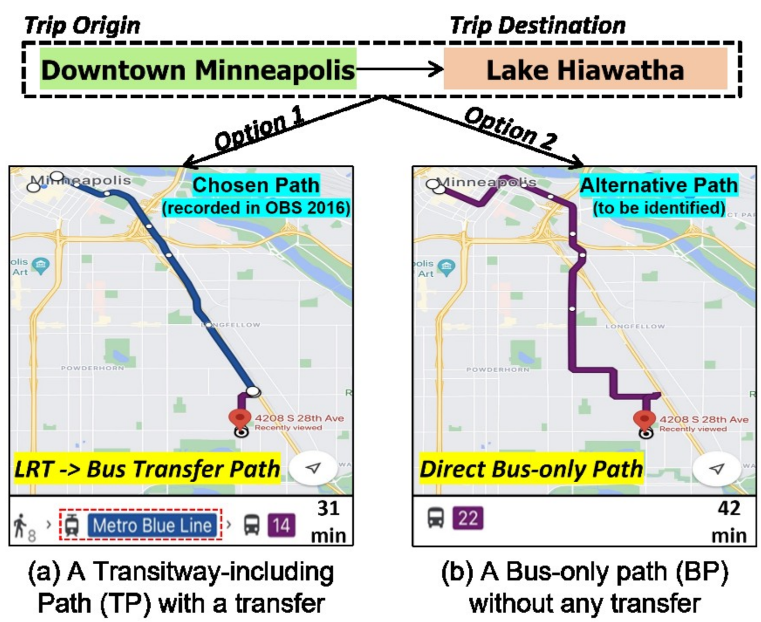


Figure 3.1 An example of a path choice set a survey respondent might have had

To reconstruct a respondent's path choice set, we require (1) an algorithm capable of identifying multiple transit paths for a given origin, destination, and time, (2) that the algorithm provides detailed path information such as walking time, waiting time, and number of transfers, and (3) confirmation that one of the identified paths aligns with the path recorded in the survey. By leveraging the identified path choice sets from multiple respondents, we can construct a model to quantify how various attributes of each path influence the transit user's decision. The ability to control time-related attributes, such as the total in-vehicle time for each path, is of particular significance. This control allows us to conduct modeling to investigate preferences regarding the inclusion of transitways in the path. To facilitate this modeling, we initially defined a transit path that includes at least one transitway as a Transitway-including Path (TP) and a path without any transitway as a Bus-only Path (BP).

3.1.1 Applying Schedule-Based Transit Shortest Pathfinding Algorithm

Unlike straightforward car journeys, transit travel comprises multiple legs with varied activities, necessitating a complex pathfinding algorithm. Travelers normally walk to transit nodes (access), endure waiting times, take a ride, and normally egress on foot. During their journey, inconvenient and uncertain transfers might happen, which incurs additional inconvenience. In order to evaluate different transit paths, researchers assign weights to each travel stage, considering differing time cost perceptions, with the additional transfer penalty (Train, 2009). That weighted sum of travel times plus the transfer penalty and monetary cost—where the fare differences are minimal in the Twin Cities, so less meaningful—is often referred to as the transit path's (generalized) cost or (dis)utility. The above transit path concepts are illustrated in **Figure 3.2**.

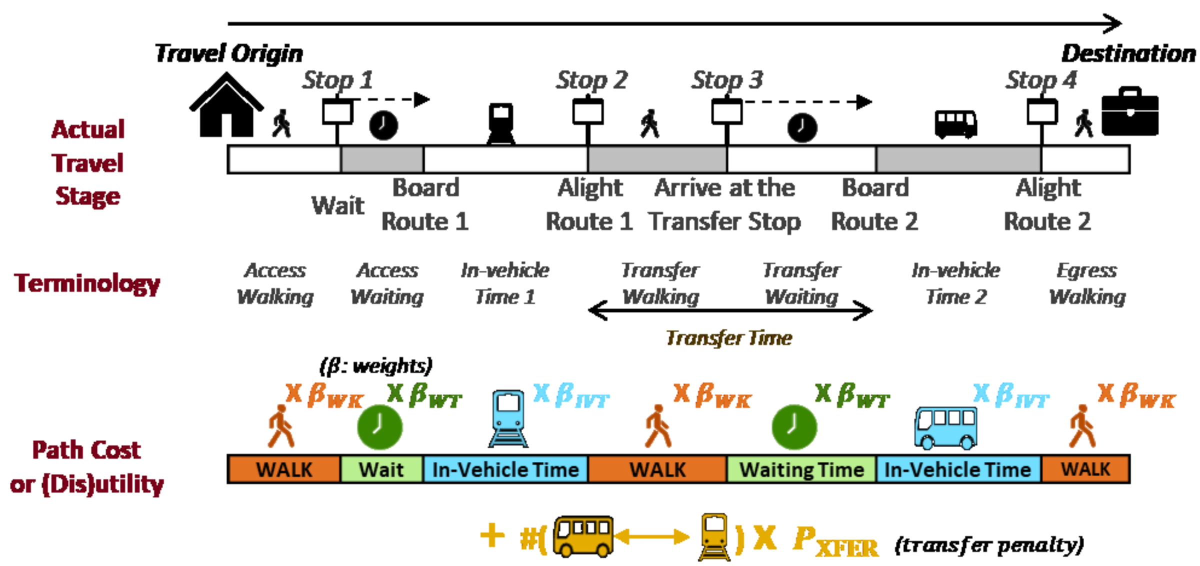


Figure 3.2 A transit path's actual travel stages and path cost

Many transit routes adjust headways based on fluctuating demand throughout the day, while transitway routes maintain consistent headways. Express bus services in the Twin Cities, mainly operating during peak times, represent an extreme case, contrasting with the steady headways of transitway services. This ensures reliable transitway service usage for eligible OD pairs during off-peak hours when conventional bus routes are limited, as indicated by the previous chapter's statistics.

However, the conventional approach of utilizing path assignment methods based on a uniform headway assumption proves inadequate when examining the impacts of transitways at different times of the day. To address this, we adopt the Schedule-Based Shortest Path (SBSP) algorithm proposed by Khani et al. (2015). Leveraging General Transit Feed Specification (GTFS) data for transit supply and the OD data from the OBS 2016 dataset, this algorithm incorporates both passenger departure times and transit routes' schedules. Drawing inspiration from Dijkstra's technique, the algorithm employs a node labeling approach to output a single "shortest" path based on a predefined objective function, such as minimizing total travel time. In addition, we integrate a trip-elimination technique proposed by Tomhave & Khani (2022) into our SBSP algorithm. For better results, the enhanced version also embeds an Open Street Map (OSM) for accurate access/egress pedestrian routing. These enhancements ensure the identification of multiple reasonable paths for a given origin/destination input. **Figure 3.3** illustrates how the enhanced SBSP algorithm works and what it can identify with the given survey input.

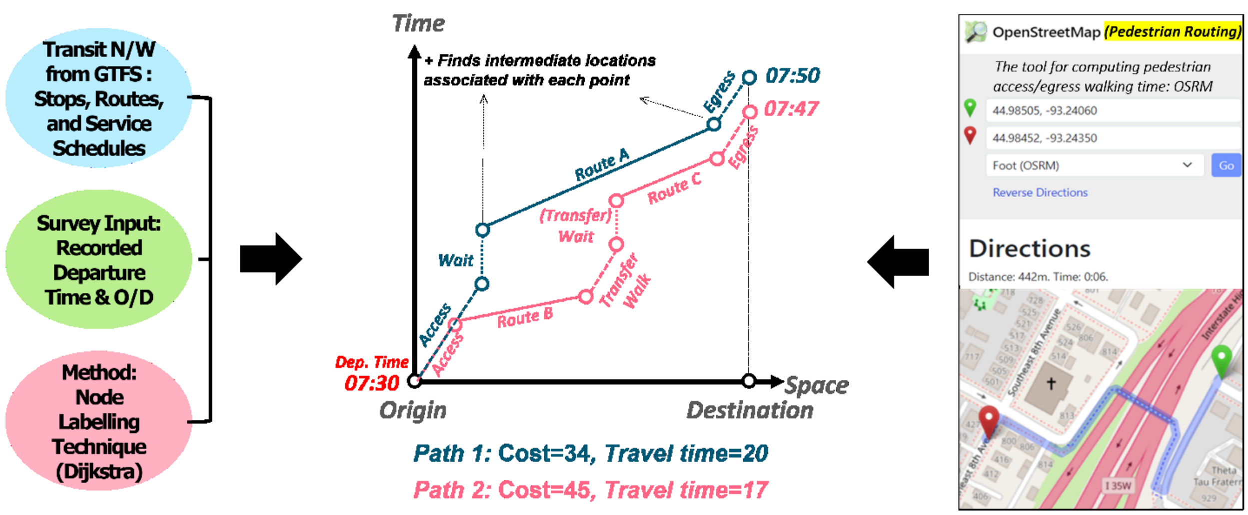


Figure 3.3 An illustration of the enhanced SBSP (Tomhave et al., 2022)

Owing to certain constraints in the algorithm, the input for this analysis was confined to a subset of data extracted from OBS 2016. Specifically, instances were selected where respondents exclusively utilized "on foot" as both their access and egress modes. Moreover, it is worth noting that the OBS 2016 data provides travel start times recorded on a one-hour basis, leading to a limitation in precisely calculating waiting times at the initial transit node of a path (access waiting time). In lieu of accurate identification, which is impossible, we opted for inferred estimates of access wait time, employing behavioral assumptions derived from Fan & Machemehl (2009). See Tomhave & Khani (2022) for more limitations of applying the SBSP algorithm to the OBS 2016 data. The pertinent weights and thresholds employed by the algorithm are succinctly outlined in **Table 3.1**, mostly following the previous work but modified if needed for better path identification. For a more tangible representation, **Figure 3.4** illustrates example path sets, each encompassing two distinct paths identified for a specific respondent.

Table 3.1 The applied SBSP algorithm's travel time weights and algorithm thresholds for pathfinding

Category	Item	Value
Weight	In-vehicle time	× 1.000 (baseline)
	Access/transfer waiting time	× 3.107
	Access walking time	× 1.746
	Egress walking time	× 2.893
	Transfer walking time	× 2.318
	Additional transfer penalty	+8.8 minutes/transfer
Threshold	Access/egress walking distance	<1.1 miles
	Transfer walking distance	<0.2 miles
	Initial/transfer waiting time	<20 minutes
	Total travel time	<5 hours

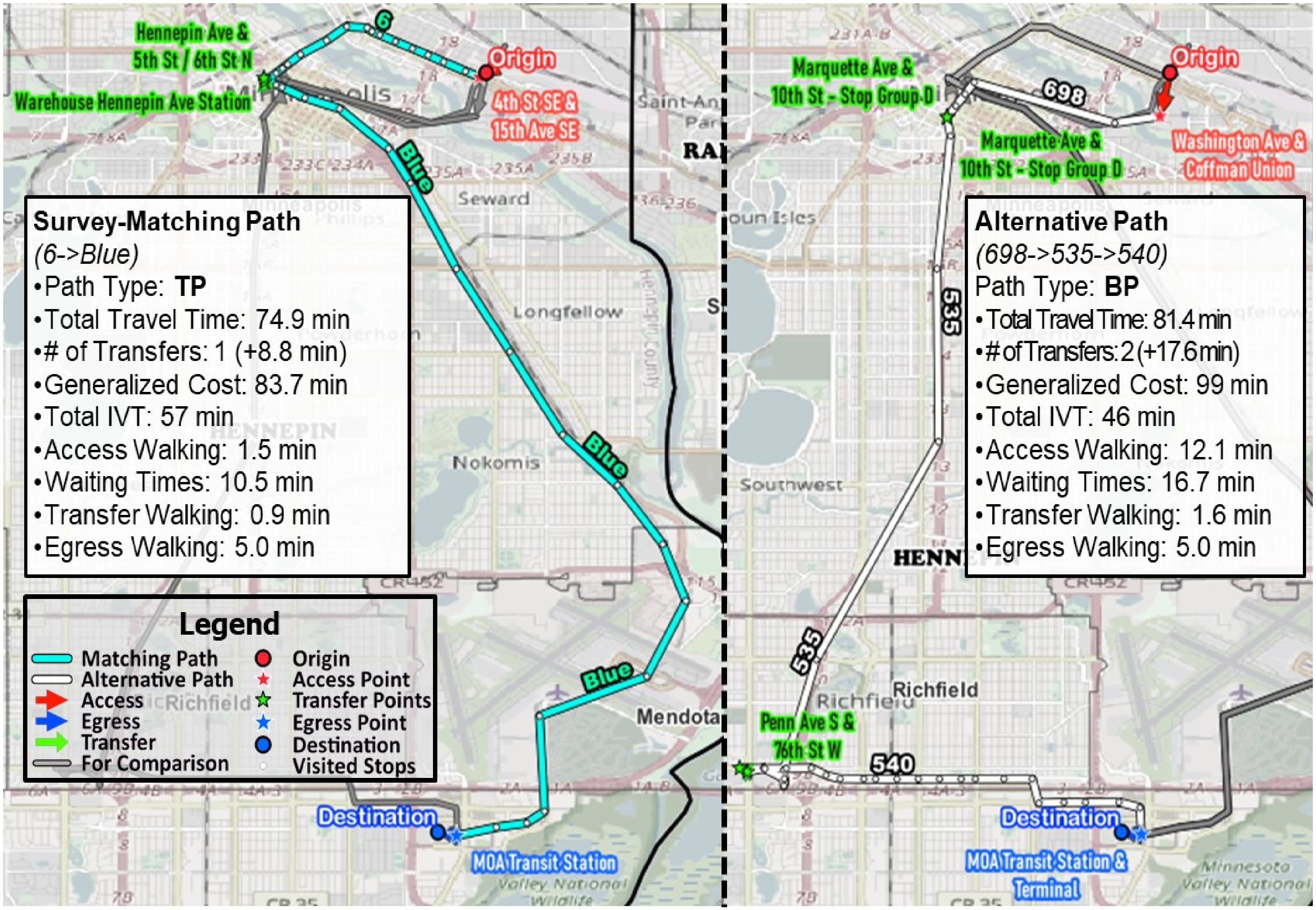
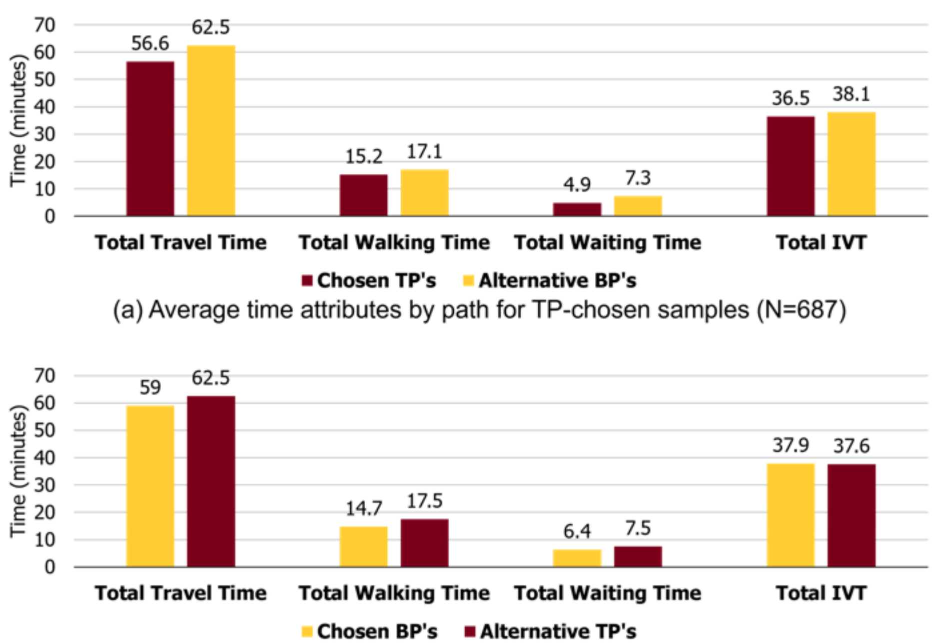


Figure 3.4 An example of the enhanced SBSP execution for a sample: two transit paths identified

3.1.2 Pairing the Identified Paths to Create the Path Choice Dataset

After employing the enhanced SBSP algorithm to identify multiple paths for each survey input, a filtering step was applied to retain samples with only realistic alternatives for comparison. Paths with generalized costs deviating by more than 50% or exceeding 20 units compared to the respondent's shortest path were excluded. This process resulted in a substantial reduction, leaving respondents primarily with a single—often their chosen—path. Further refinement included keeping respondents with a choice set comprising at least one TP and one BP option.

This collection, named "the path choice dataset," serves as our focus for TP-BP comparisons. Within this dataset, 687 respondents opted for TP, while the remaining 420 chose BP. **Figure 3.5** illustrates the average travel time components for the chosen path and the alternative path identified by the algorithm. The statistics summarized in the upper panel indicate, among the TP-selected respondents, how the average path attributes differed between their used (or survey-matched) TP and the inferred alternative BP. The TP-used passengers had an average total travel time of 56.6 minutes for their TP choice. It is smaller but not much different than their alternative BP's value, 62.5 minutes. Likewise, other average path attribute values were not dramatically different between the path options. On the other hand, the lower panel allows us to observe similar trends for the BP-used group. The statistics collectively reconfirm that the used path's attributes are smaller than the alternative but not too much. These findings suggest that our method effectively identified and paired the only alternative path that passengers would realistically consider choosing.



(a) Average time attributes by path for TP-chosen samples (N=687)

(b) Average time attributes by path for BP-chosen samples (N=420)

Figure 3.5 Average time attributes of the path choice dataset by their recorded path choice

To explore passengers' path choice behavior related to transitway inclusion and transfers, we categorized them into four groups (see **Figure 3.6**). Group 1 includes passengers with direct transit paths in both their TP and BP. Group 2 comprises passengers with transfers in both paths. Group 3 involves passengers with a direct TP and a transfer-including BP, while Group 4 exhibits the opposite, featuring a direct BP and a TP with transfers. Each path's number and order of transfers do not affect the grouping.

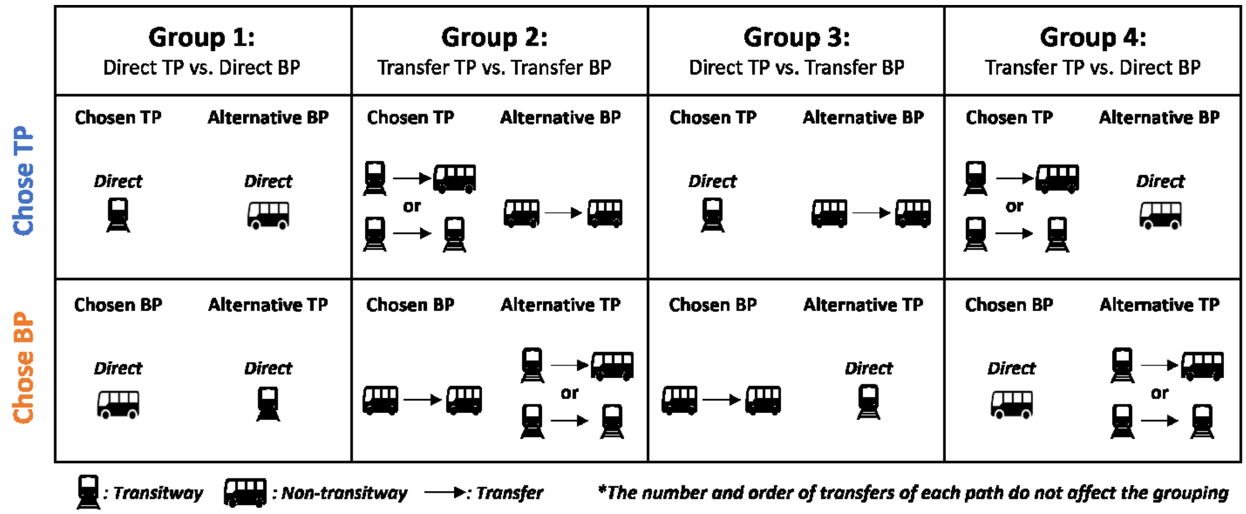


Figure 3.6 Passenger grouping depending on their paths' transfer inclusion

Figure 3.7 visually represents the distribution of path choices within each group. At an aggregate level, when individuals had comparable TP and BP options, 62% favored the TP. Notably, except Group 4, every group exhibited choice proportions exceeding two-thirds. Although the TP choice proportions for Group 3 (97%) and Group 4 (23%) may seem extreme, they are plausible, as one path for these groups is less burdensome than the other due to the presence of a transfer in only one of the options. It is particularly noteworthy that 23% of individuals in Group 4 opted for the transfer-included TP over the direct BP. Among the four groups, Groups 1 and 2 have a controlled number of transfers between the two options, ensuring a more equitable comparison. This collectively confirms that passengers generally prefer TP to BP, especially when the transfer specifications between the two paths are similar. Additionally, the data suggests that the inclusion of a transitway sometimes surpasses the inconvenience associated with a transfer. However, it is uncommon for individuals to choose the BP with a transfer over a direct TP.

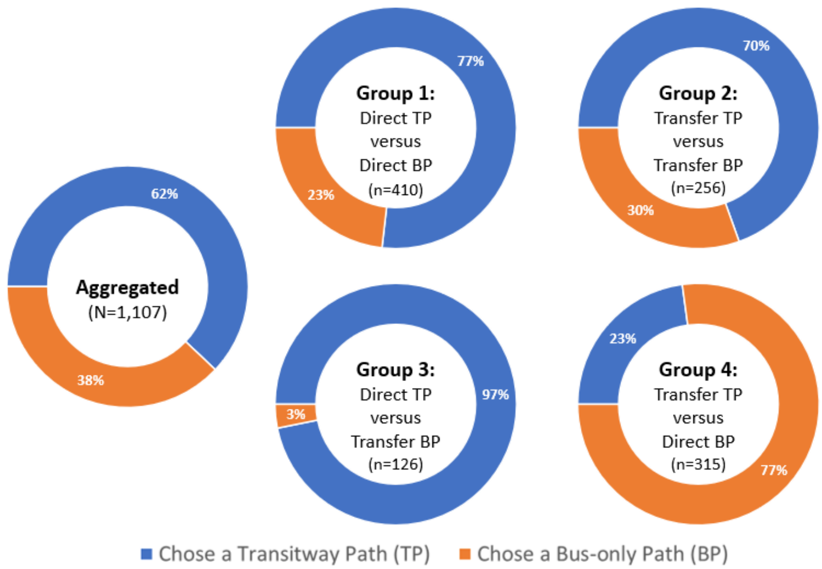


Figure 3.7 Path choice proportion of the paired path choice dataset by group

3.2 LOGISTIC REGRESSION ANALYSIS

This section aims to investigate how passengers value or perceive their TP option over the BP depending on the comparative path attributes. To answer this research question, we estimated logistic regression models and analyzed the results separately based on passengers' demographics and travel contexts. Specifically, we seek to identify (1) the general preference for TP over BP when both options have the same travel time attributes and (2) the extent to which passengers consider TP and BP to be perceived equivalent.

Logistic regression is a statistical model that uses a logistic function to obtain how much probability the value of an independent variable will lead to having a specific value of a binary dependent variable. A univariate logistic regression has a mathematical form like the following equation, where the independent x value is assumed to affect the probability of the dependent variable y being 1, which has λ (intercept) and μ (coefficient) as parameters to be estimated.

$$p(y=1) = \frac{1}{1+\exp(-\lambda(x-\mu))}$$

Our binary dependent variable, *Path Choice*, was generated depending on the path the passengers used. If they used the TP option, it was coded as one ($y = 1$), whereas zero ($y = 0$) was assigned to BP users. The independent variables were derived as "difference variables" that measured the differences between the attributes of the two path options for each passenger, namely differences in total travel time, in-vehicle time, and walking time. Specifically, we calculated those difference variables by subtracting a time attribute of a passenger's TP option from their BP counterpart's value. The resulting values indicate the relative advantage of the TP option over BP. For instance, a variable *Total Travel Time Difference* of 15 for a passenger implies that their TP option was 15 minutes shorter than their BP. In contrast, the value indicating -20 suggests that a passenger's TP option was 20 minutes longer than their BP.

We estimated multiple univariate logistic regression models instead of one multivariate model to measure the TP's attractiveness on the path choices conditional on the time differences. The reasons are to estimate high-level quantities representing a system or demographic group-wise perception difference, and univariate models are more suitable for visualization. **Table 3.2** summarizes the results of three univariate regression models, each representing a different difference variable. Waiting times were not made as a difference variable as they are less controllable by passengers, and the survey data's time of departure resolution did not allow the algorithm to measure it precisely, as introduced in the previous section.

Table 3.2 The univariate logistic regression models result

Model	Model Parameter	Estimate	Standard Error	Z-Value	P-Value
Model 1: Travel Time (min)	Intercept	0.416	0.066	6.286	<0.001
	Difference Coefficient	0.062	0.006	10.982	<0.001
Model 2: In-vehicle Time (min)	Intercept	0.485	0.062	7.797	<0.001
	Difference Coefficient	0.008	0.005	1.611	0.107
Model 3: Walking Time (min)	Intercept	0.515	0.064	8.022	<0.001
	Difference Coefficient	0.056	0.007	7.797	<0.001

The logistic formula underlying each model suggests positive intercepts would indicate people's general preference for TP to BP when they have the same time attribute. Likewise, positive variable coefficients mean an increase in the difference variable (i.e., one additional minute of time-saving realized by choosing TP over BP), which would lead to a higher probability of choosing the TP. All models had statistically significant (P-value<0.05) intercept estimates. These findings confirm that people prefer TP when the travel times are the same. This appears to be affected by features of TP, such as reliability-related factors like on-time performance and even headways, which also contribute to transfer convenience. However, only two out of the three models, namely total travel time and walking time differences, had significant coefficient estimates. In other words, the insignificant coefficient of the *In-vehicle Time Model* (Model 2) implies that it lacked evidence of passengers weighing on incremental in-vehicle time savings between TP and BP.

Figure 3.8 displays Models 1 and 3's resultant logistic curves of the TP choice probability over BP (y-axis) along the time difference being examined on the x-axis. Two auxiliary lines were drawn at $x = 0$ and $y = 50\%$ to obtain numbers with special meanings. The intersection of the former and the curve represents the conditional choice probability when TP and BP have the same attribute value or their difference is zero. On the other hand, the intersection of the latter and the curve represents how much path attribute difference, on average, is required for TP and BP to be considered equivalent. The intersections reveal that if the TP and BP options have the same total travel time, around 60% of passengers would choose TP, and if they have the same walking time, about 63% would. Additionally, passengers would perceive the BP option similarly attractive if it were 6.7 minutes shorter in total travel time or 9.3 minutes shorter in walking time than the TP option.

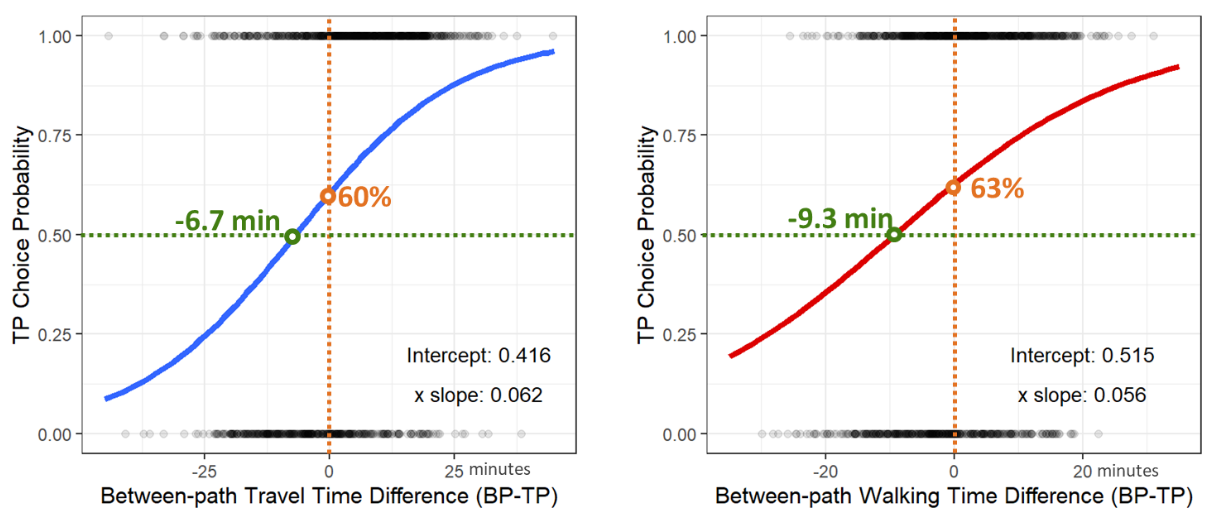
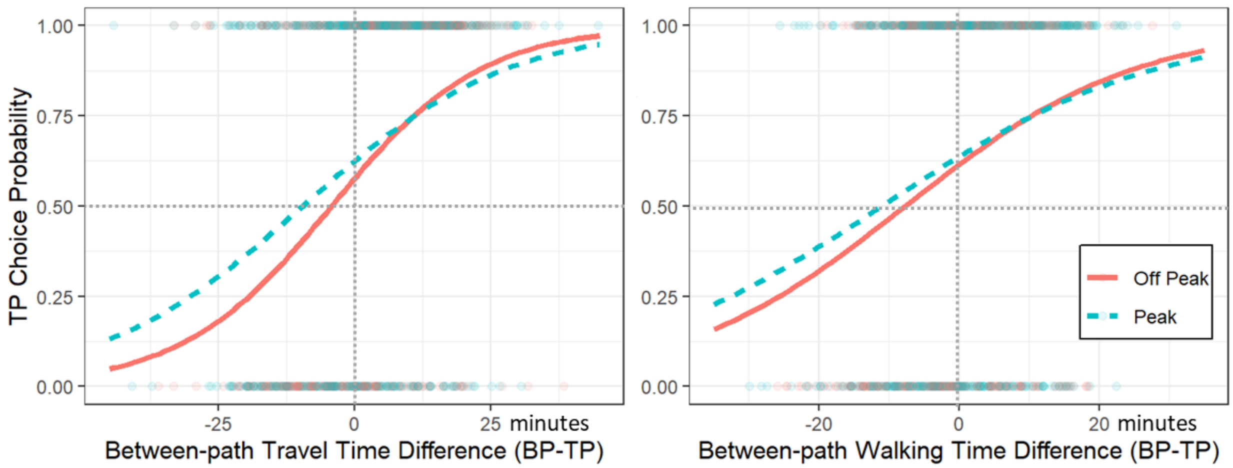
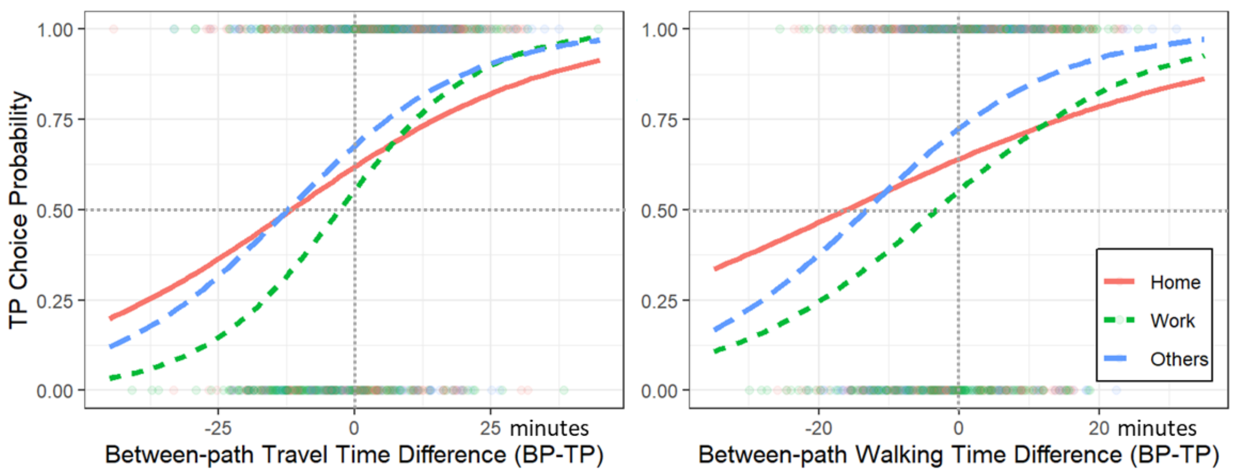


Figure 3.8 Logistic regression curves for the total travel time and walking time differences

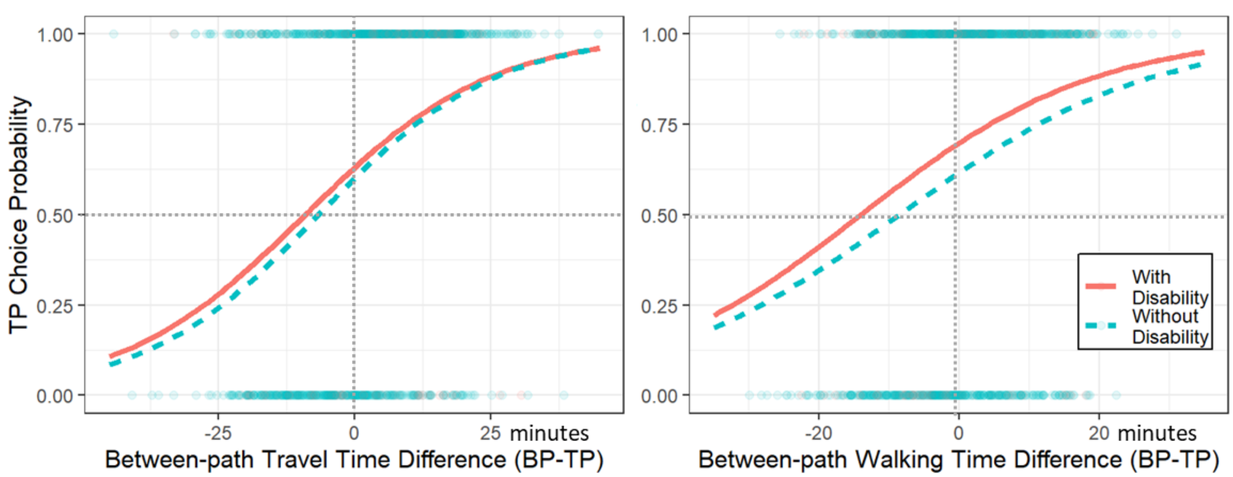
Unlike the above curves drawn using all the paired paths, we could separately estimate the models and draw associated curves by travel contexts/demographic groups to which each passenger belonged. The division of the data is conducted similarly to the mosaic plots in the previous chapter but without the regional context due to the smaller sample size. **Figure 3.9** illustrates those results. The first panel shows the separately fitted logistic curves for off-peak travelers and peak time travelers; the second the destination type of tip purpose (where "work" represents the commute trips); and the last panel by disability.



(a) Logistic regression curves: time of departure



(b) Logistic regression curves: destination type



(c) Logistic regression curves: disability

Figure 3.9 Disaggregated logistic regression curves by travel context and demographic

All the intersections were located in the upper-left quadrant of each graph, reconfirming the transit users' TP preference. Notable observations from each panel of the above figure are summarized as follows:

a) Peak-time travelers are more likely to choose TP than off-peak-time travelers, particularly for choice situations when the total travel time of TP is longer than that of BP (x<0 region).
b) When TP and BP travel times are similar, the path choice probability for work destinations (commutes), compared to the others, was closest to 50%. This implies commuters are more sensitive to travel/walking time savings *per se* than other TP advantages.
c) For the walking time difference, the "people with disability" curve is located above the "people without disability" curves, indicating the group's strong TP preference.

This chapter delved into discerning user preferences for transitway paths by utilizing the choice sets identified from the survey. It is crucial to acknowledge certain limitations in our approach. Firstly, the method employed for comparing paths between TP and BP did not account for respondents utilizing non-walk access/egress modes, such as park-and-ride, thereby restricting the generalizability of our findings. Secondly, a noteworthy proportion of short-distance trips (as depicted in **Figure 2.11**) were excluded from the TP/BP comparison analysis since many such trips lacked reasonable alternatives and were not paired for comparison. Consequently, our results exhibit a bias toward long transit trips, as shown in the average values of **Figure 3.5**, whereas, in practice, hour-long transit trips are not the average trip. The logistic regression results and proposed representative numbers are, therefore, more applicable to long-distance trips. Lastly, there is room for reformulating the definition of TP and BP to align with people's actual perceptions, for instance, incorporating criteria such as distance or time covered by the transitway portion of the path.

CHAPTER 4: OFF-PEAK TRANSIT DEMAND ANALYSIS WITH AFC

Many transit operators have embraced the Automatic Fare Collection (AFC) system, utilizing stored-value transit cards or QR codes in smartphone apps. This system not only enhances passengers' convenience in fare payment but also generates extensive transaction data, offering valuable insights for transportation studies. We used the Twin Cities' long-term AFC data to conduct an off-peak transit demand analysis, focusing on the transfer, commutes, and the COVID-19 pandemic.

The AFC data we utilized originates from taps on transit cards, such as the Go-To card and UMN's U-card, representing a significant proportion of Twin Cities transit users both spatially and temporally. The dataset provides continuous transit ridership records, surpassing traditional surveys' capacity in terms of volume and timestamp accuracy. Leveraging the exhaustiveness of AFC data compensates for the limited representativeness of OBS 2016 used in the previous chapters. We used the AFC dataset from January 1, 2018, to August 28, 2023, excluding some days during the early stage of the COVID-19 pandemic when the fares were not collected. It features 146,211,190 total transactions, or 72,167 transactions per day. **Table 4.1** presents a comparative analysis of AFC and OBS 2016 observations.

Table 4.1 Datasets' count and representativeness comparison: OBS 2016 and AFC

Category/Data Source		OBS 2016		AFC Data	
Data Collection Period		Apr 2016 – Feb 2017		Jan 2018 – Aug 2023	
Total Count		30,491		146,211,190	
Sub-category and their value		Count	Proportion	Count	Proportion
Time of Day	Early Morning (4 AM – 6 AM)	117	0.4%	3,980,560	2.7%
	AM Peak (6 AM – 9 AM)	8,236	27.0%	35,341,095	24.2%
	Mid-day (9 AM – 3 PM)	8,895	29.2%	42,494,854	29.1%
	PM Peak (3 PM – 6:30 PM)	9,675	31.7%	47,558,719	32.5%
	Night (6:30 PM – 4 AM)	3,568	11.7%	16,835,962	11.5%
	Monday	1,150	3.8%	24,668,947	16.9%

Day of Week	Tuesday	8,713	28.6%	26,847,149	18.4%
	Wednesday	9,518	31.2%	26,734,388	18.3%
	Thursday	10,721	35.2%	25,949,354	17.7%
	Friday	389	1.3%	23,398,700	16.0%
	Saturday	0	0.0%	10,678,991	7.3%
	Sunday	0	0.0%	7,933,661	5.4%
Season (For AFC, Jul 2018 – Jul 2023)	Spring (Mar – May)	7,174	23.5%	26,135,507	21.7%
	Summer (Jun – Aug)	3,902	12.8%	27,265,989	22.7%
	Fall (Sep – Nov)	12,885	42.3%	36,430,583	30.3%
	Winter (Dec – Feb)	6,530	21.4%	30,512,984	25.4%

Recognizing the inherent limitations of the survey data, especially in representing specific times, we capitalize on the continuous nature of AFC data. This approach of data complementation, suggested by Chapleau et al. (2018), fills time-related representativeness gaps in survey datasets. The continuous nature of the AFC is illustrated in **Figure 4.1**. The notable long-term ridership fluctuations in the figure highlight the impact of the COVID-19 pandemic on transit usage patterns.

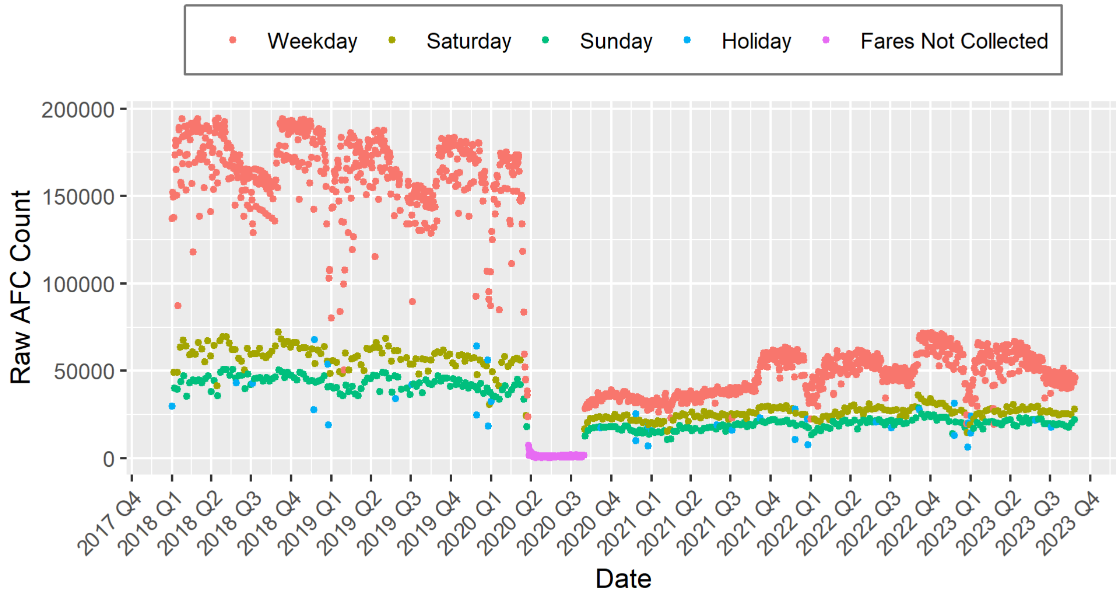


Figure 4.1 Raw AFC data count by day

Raw AFC data, while extensive, needs more detailed information on transfers and trip purposes for conducting the travel behavior analysis. To address this, we employ the "trip chaining" technique, inferring either the boarding or alighting side of each trip, which is missing in the raw data. Despite the enriched AFC dataset after trip chaining, limitations persist in capturing detailed travel contexts. After, we introduce survey-trained ML models to complement trip chain records, addressing gaps in travel behavior information. This section outlines these models' development, training, and selection process to make reasonable inferences on the AFC dataset. Finally, the last section presents visually compelling results incorporating ML-inferred trip purpose attributes. This demonstrates changes in off-peak commute and transfer behavior in the Twin Cities between 2018 and 2023.

4.1 TRIP CHAINING

As the name implies, the trip chaining technique interconnects multiple trips recorded in the raw AFC data. This interconnection is imperative due to the Twin Cities' AFC system, which, in most cases, mandates passengers to tap their cards only upon boarding. To infer missing alighting locations, trip chaining assumes that subsequent taps from the same card should be near the previous trip's alighting location. **Figure 4.2** illustrates what we can expect from the process using a hypothetical travel diary for a day. Assume Trips 1, 2, and 4 are transit trips recorded in the AFC dataset, while Trip 3 involves an out-of-system transport mode (e.g., electric scooter sharing). The absence of Trip 3 in the AFC data poses challenges in tracing the travel trajectory, preventing the inference of alighting/transfer for Trip 2 and transfers between Trips 2 to 3 and Trips 3 to 4.

The chaining step leverages partial information stored in an AFC record: we can access on which bus, heading in which direction, the tap was recorded. By merging General Transit Feed Specification (GTFS) data, containing detailed transit operation schedules, with the AFC data, we can identify each tap's most plausible alighting stop. Determining the most plausible stop and further technical details are available in Kumar et al. (2018). It is worth noting that any errors in GTFS and/or AFC that block the matching of a GTFS vehicle trip (the arrows) to an AFC transaction would cause the same impact of our dataset having the missing records as described in Trip 3 of the example.

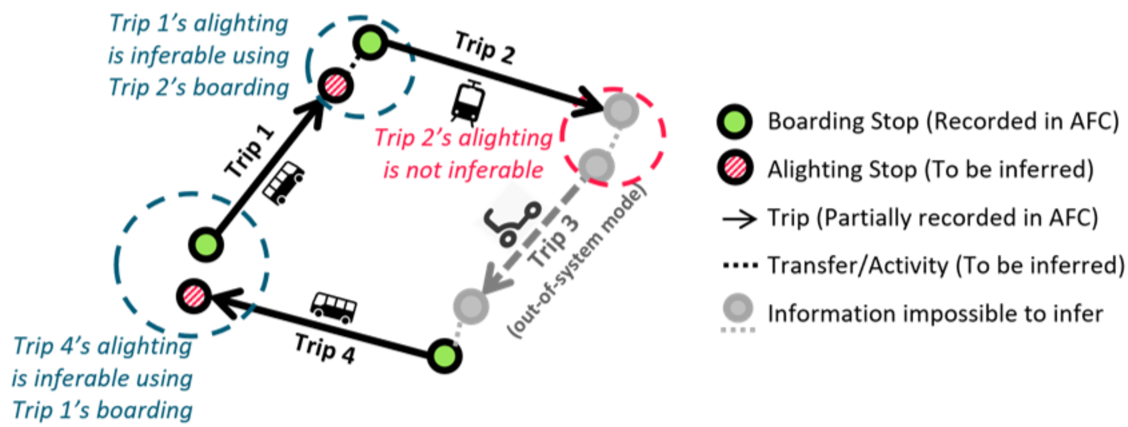


Figure 4.2 What trip chaining algorithms can infer

The trip chaining algorithm was applied to the raw AFC data, successfully appended alighting stop information to 73,058,936 transactions, and identified 14,475,138 transfers. These "practical" transfers indicate sequential boarding occurring within 30 minutes after the previous alighting, offering richer information than nominal transfers that the AFC system uses—any sequential boarding taps happened within 2.5 hours. Additionally, transfers attached to the trip chain data stores transfer origin and destination and the time elapsed for the transfer activity, which are also unavailable in the survey.

However, data loss is inherent in the process. Reasons for data loss encompass AFC-side errors, GTFS-side discrepancies, and passenger behaviors. AFC-side errors include missing stop and route information, which is crucial for matching a GTFS vehicle trip to it. The GTFS-side limitations stem from discrepancies in delayed/faulty updates of the published schedules and unexpected events that happen in real time, including traffic congestion. This is mainly observed in the early stage of the COVID-19 pandemic and the periods when the new major transitway routes (i.e., C/Orange/D lines) are launched. Moreover, passenger behaviors, such as intentional or mistaken boarding without card tapping and utilization of other modes of transportation for some trip legs, contribute to data quality challenges. Also, there are many cases where passengers mistakenly tap multiple times on the same fare collector within a short interval, and many cardholders made only a single trip—making the inference difficult or impossible. Finally, an algorithm-related data loss happens because increasing the tolerance in the changing steps lowers the output quality. If we allow the algorithm to exploit more usable information with loose tolerances, it will chain more trips, but the results' accuracy becomes less reliable.

Therefore, we evaluated and tracked the "performance" of our trip-chaining algorithm by how much it successfully exploited usable information to chain transactions, leaving the transactions impossible to handle out of the count. The following equation describes its derivation. The term T is the number of chained transactions, G is the count of AFC transactions to which at least one GTFS vehicle trip was matched—so they are the only eligible transactions for the chaining. Next, D is the number of additional duplicated taps at the same stop or fare collector that occurred within 5 minutes, and R is the total raw transaction. In the algorithm, duplicated taps are first discarded. Afterward, if we cannot match a transaction to a GTFS trip, not only the very transaction is discarded, but it also limits the inference on the previous trip as well, leading the data amounted to $(R - D - G)$ to become unusable. To balance the data loss and trip-chaining accuracy trade-off, we heuristically adjusted the algorithm criteria to achieve 80% performance on average.

$$Data\ Retrieval\ Performance = \frac{T}{G - (R - D - G)}$$

Figure 4.3 illustrates the algorithm's data retrieval performance tracked daily, excluding the period between Mar 26, 2020, and Jul 31, 2020, characterized by a daily raw AFC count below 10,000 taps when the fare payment was optional. Numerous data quality issues during this period prompted its exclusion. Visible changes in performance were observed before and after the outbreak of the COVID-19 pandemic, coinciding with the format changes of Metro Transit's GTFS in the summer of 2020.

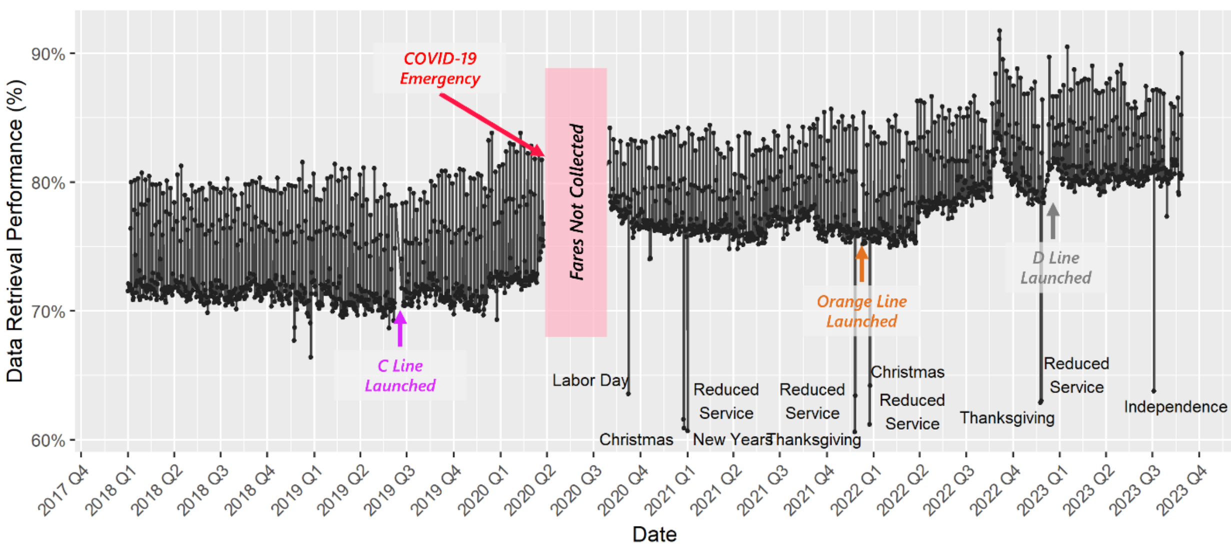


Figure 4.3 Data retrieval performance tracked for applying the trip chain algorithm

With the enriched trip chain data after the chaining, transfers are classified based on whether their transfer origin or destination involves transitway stations (**Figure 4.4**). We focused on transitway transfers because they are designed to facilitate transfers, as emphasized in the transit plan of the region—*In transitways where different service types operate in conjunction with each other, these services should be coordinated to facilitate convenient and reliable transfers* (Metropolitan Council, 2016).

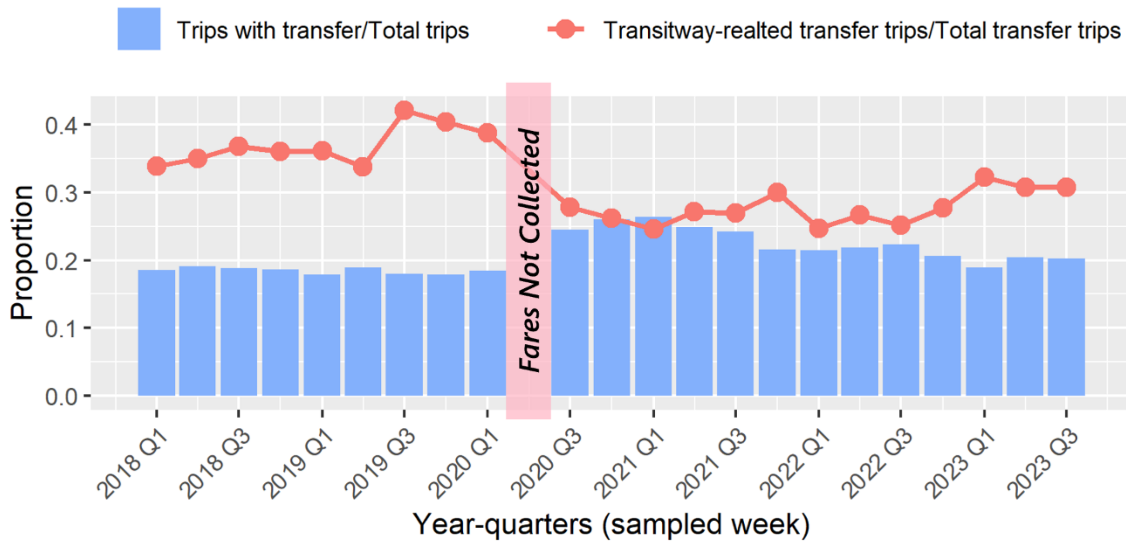


Figure 4.4 Transfer trip proportion and the share of transitways in transfer trips

The blue bars represent the proportion of transfer trips computed by dividing the number of trips with transfers by the total number of linked or chained trips. The red line depicts the share of transitway-related transfers, indicating transfers involving transitway stops divided by the total number of transfer trips. The proportion of transfers increased after the COVID-19 outbreak. However, the share of transitway-related transfers in the total transfer trip chains decreased for the same period above. Other observations include:

- Approximately 21% of transit travels involved at least one transfer.
- About 32% of transit transfers were associated with transitways.
- The increase in transfer proportion from 2020 3Q could be due to frequent changes in the service during the pandemic, reducing the number of destinations reachable without transfer.
- Transitways remained relatively reliable during the pandemic, as the decreased transitway transfer share indicates that the increase in the transfer rate mainly occurred between buses.

4.2 TRIP PURPOSE INFERENCE USING MACHINE LEARNING

The trip chain data can be further enhanced by integrating it with the survey, facilitated by artificial intelligence-powered machine learning (ML) techniques, which augments the dataset by incorporating essential features for behavioral analysis.

4.2.1 Training and Selecting Machine Learning Algorithms

Travel inherently serves social and economic activities, and discerning the implicit travel purposes associated with observed trips is essential for comprehensive transit planning and ridership diagnosis. While trip chains still lack explicit information on trip purposes, OBS 2016 contains passengers' explicit inquiries about trip purposes. By extracting correlations between passengers' trip purposes and revealing spatiotemporal travel context from the survey, we aim to transfer this knowledge to the trip chain data to augment its richness. This entails developing a statistical model capable of predicting a passenger's trip purpose based on new input variables extracted from a trip chain record.

Past efforts to augment AFC trip chain results with trip purpose included rule-based assignment (Lee & Hickman, 2014) and inference from various data sources (Alsger et al., 2018). In contrast, we leverage recent advancements in hardware capabilities to train machine learning (ML) models for similar purposes. Our models are trained to label survey data with destinations categorized as home, work, school, social, shopping, or other purposes. The ML models utilize various input variables, including travel time-related features (as introduced in **Table 4.1**), geographic locations based on the Transit Market Area (Metropolitan Council, 2015) with some modifications/additions, type of service (BRT/LRT/express bus) used, and the inclusion of transfer(s). These predictors are present or can be inferred in both the survey and trip chain domains, enabling the transfer of ML models trained from the survey to enhance trip chain data.

Figure 4.5 visually represents the application of pre-trained ML models to the trip chain data. The left-hand-side dots represent the original trip chain results (their destination points) whose trip purpose is unknown and thus colored black, but they store some information that could also be found in the survey. The survey-trained ML models predict or label each dot one of the trip purposes available in the survey. Therefore, we can plot our final data with trip purpose shown in the right-hand-side maps, where the results of the two example ML models' are color-coded to distinguish different purposes attached to each trip chain.

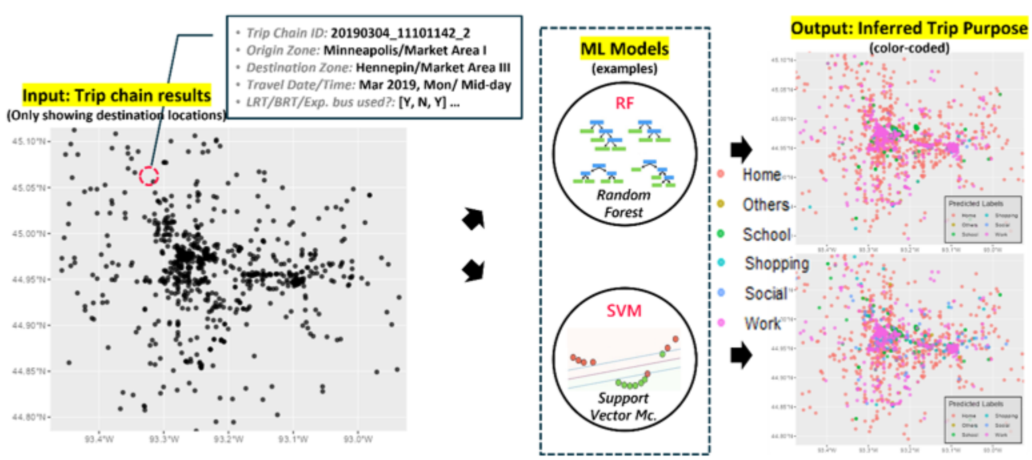


Figure 4.5 Illustration of the ML application: trip purpose augmentation to the trip chain results

Numerous ML models exist for such a data-labeling task or supervised classification problem. Koushik et al. (2020) summarized those models and how different transportation studies used different models. Rigorous testing was conducted to assess each model's accuracy for overall and "Work" label predictions, comparing the labeling results with the true trip purpose distribution recorded in the survey. **Table 4.2** summarizes the evaluation results, where the last column highlights the absolute mean percentage errors for surveyed work trips and predicted work trips in terms of destination (Downtown St. Paul/Minneapolis) and time (AM peak/others). While all models demonstrate similar accuracies, the neural network outperformed in terms of similarity to survey data, validating its efficacy for trip purpose inference in subsequent sections of our analysis.

Table 4.2 The validity evaluation of the trained machine learning models

Machine Learning Model Name	Overall Accuracy (1-test error)	"Work" Label Accuracy	"Home" to "Work" Label Counts Ratio	Spatiotemporal Discrepancy
Survey (reference)	**NA**	**NA**	**1.252**	**0%**
Random Forest	65.3%	86.6%	1.065	4.53%
Adaboost	64.4%	86.9%	1.488	6.43%
Gradient Boost	64.5%	87.6%	1.002	7.70%
XG Boost	62.6%	82.1%	1.072	4.80%
Neural Network	**65.8%**	**89.2%**	**1.280**	**2.98%**
Support Vector Machine (Linear)	64.8%	89.4%	0.594	5.10%
Support Vector Machine (Radial)	63.9%	89.1%	0.882	5.90%

Note: The last two columns indicate how similar each model's prediction distribution is to the survey data. The closer the label counts ratio is to the reference value of 1.252, and the closer the discrepancy to 0%, the better the result.

4.2.2 Commute Trip Analysis from the Neural Network Model's Inference Result

For result presentations, we defined the three periods in response to the COVID-19 pandemic. Specifically, the pre-pandemic period is defined as days before the state emergency was declared (March 13, 2020; Emergency Executive Order 20-01) and the post-pandemic as days after the major daily restriction was lifted (May 14, 2021; Emergency Executive Order 21-23), and defined any days between the two as a pandemic days.

Figure 4.6 shows the variations in AM peak and non-AM-peak transfer trip proportions for each destination area—Downtowns for Downtown Minneapolis and Downtown St. Paul combined—and the abovementioned period. Notably, when the destination is downtown, AM-peak transfer trip proportions are consistently lower than those during non-AM-peak hours. During the pandemic, irrespective of the destination, transfer trip proportions reached their highest, followed by the post-pandemic period. This aligns with the trends observed in **Figure 4.4**, indicating that express bus services, primarily catering to the two downtowns during AM peak hours, historically facilitated suburban travelers reaching downtown without transfers but later became unavailable.

For destinations outside the downtown area where express bus services are limited, more transfers are required compared to the downtown-destined trips for every period. It is reasonable to infer that these journeys are undertaken when deemed sufficiently valuable, surpassing the transfer penalty. Specifically, during morning commutes to non-downtown locales, passengers are more inclined to embrace trips involving transfers. Conversely, non-commute excursions to non-downtown areas, primarily occurring during off-peak times, are typically undertaken exclusively by individuals who can avail themselves of a direct transit route.

Figure 4.6 Transfer trip proportions by destination location and time of day across periods

Similarly, **Figure 4.7** delves into the proportion of commute trips inferred by the neural network model with the destination labeled "Work." Regardless of the time of day, downtown areas consistently exhibited a higher proportion of commute trip destinations compared to non-downtown areas. Intriguingly, commute trip proportions at non-AM-peak saw a significant increase during and after the pandemic compared to the pre-pandemic period. This inferred shift could be attributed to heightened flexibility in work arrangements and altered commuter behavior emerging during and after the pandemic.

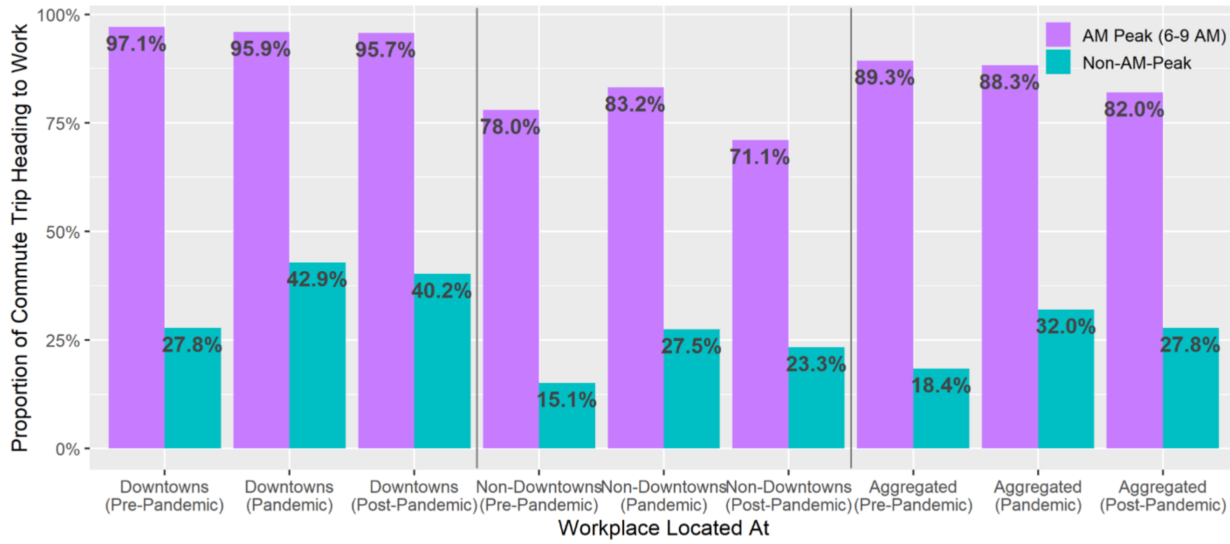


Figure 4.7 The proportion of inferred commute trips from total by workplace location and time of day

For more detailed spatial analysis, we employed three-square-mile grids covering the entirety of the Twin Cities region. Each grid is color-coded based on aggregated statistics observed from the results, with grids having counts less than 30 daily counts excluded for minimum representativeness. **Figure 4.8** exemplifies the count of trips, depicting both the origin (upper panel) and destination (lower panel). Concentrations of workplaces or commute destinations, particularly in downtown areas, resulted in lighter colors for the suburban grids in the lower panel. In contrast, commute origin (primarily residential areas) exhibited broader coverage.

Subsequent visualizations utilize conceptual representations and paneled plots to enhance clarity without multiple cartographic annotations and background maps. Each panel is designed to represent the period (pre-pandemic, pandemic, and post-pandemic) and time of day (AM peak and others). Note that the visualizations focus solely on aggregated statistics by the destination location, as in **Figure 4.6** and **Figure 4.7**, excluding information about trip origins. This deliberate choice is based on our determination that if a linked trip's destination is labeled as "Work," it inherently represents a commute trip to one's workplace. Consequently, we distinguish between more conventional peak-time commutes occurring during the AM peak and other commute trips outside the AM peak, such as those during the afternoon. For instance, an individual heading to work at 4:00 PM is considered to be making a commute but not a peak-time commute, despite the technical classification of 4:00 PM as PM peak due to the volume of passenger flows. This strategic decision allows for a clear comparison between off-peak and peak commute characteristics and facilitates an exploration of how these patterns have evolved with the onset of the pandemic.

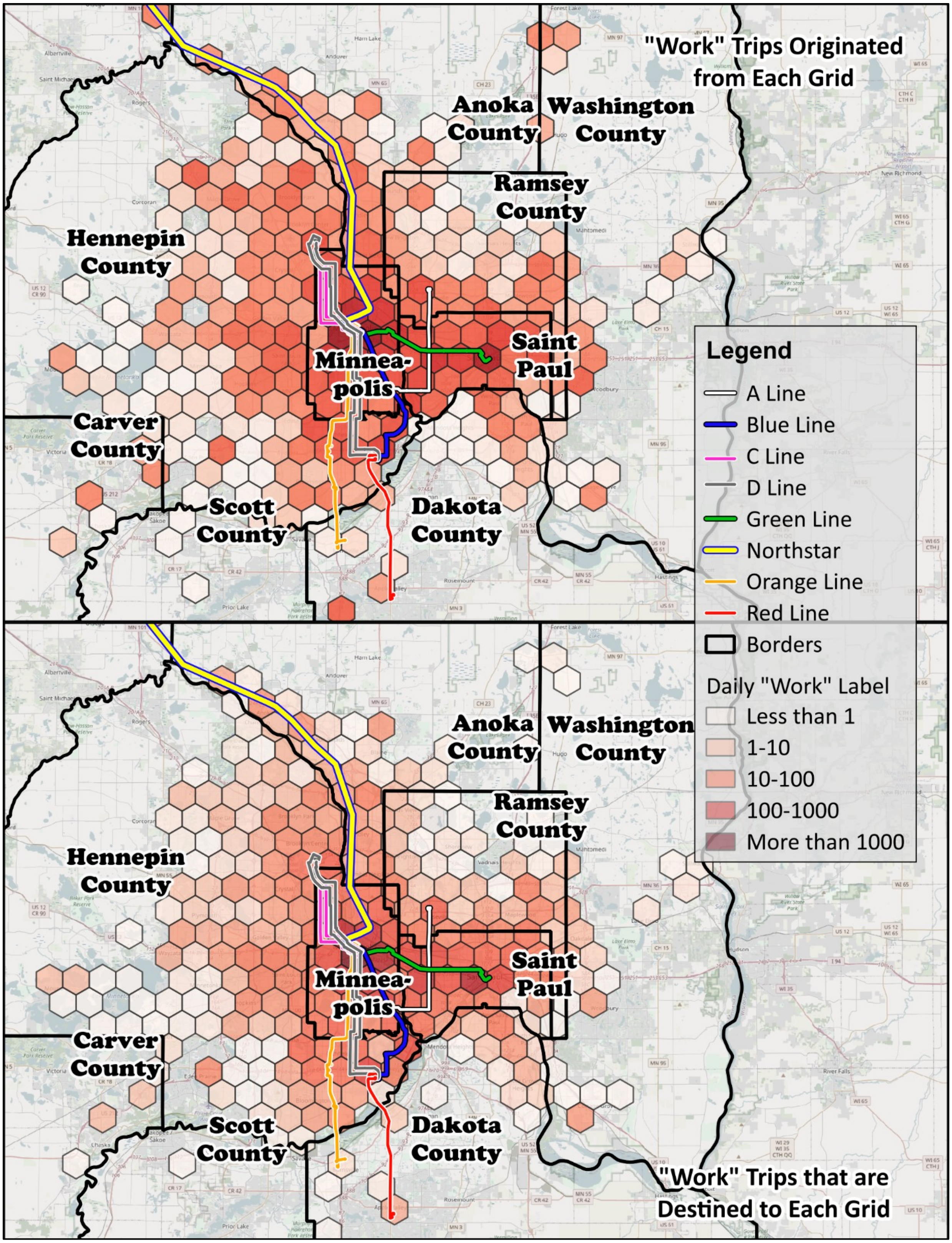


Figure 4.8 The daily average inferred counts of commute trips from the neural network model

Figure 4.9 illustrates the proportion of each grid's "Work" label or commute trips. In the first row (AM Peak), darker colors indicate a higher concentration of "Work" labeled trips during the morning peak hours, predominantly signifying traditional peak-time commutes to workplaces. The second row (Non-AM-Peak) displays significantly lighter colors, suggesting lower proportions of commute trips. This aligns with the understanding that commutes during non-AM-peak are less frequent. Comparing horizontally, there is a noticeable change post-pandemic. Simultaneously, the Non-AM-Peak's pandemic and post-pandemic panels show darker colors, suggesting increased off-peak commutes. This reconfirms workers' increased flexibility in commuting after the pandemic. Interestingly, the plots on the second column show the darkest colors among their rows. These can represent the concentration of essential workers during the pandemic since non-essential workers had more options to work from home. Alternatively, it could also represent closures of retail businesses and limited social activities allowed during the pandemic when we compare AM Peak and Non-AM Peak.

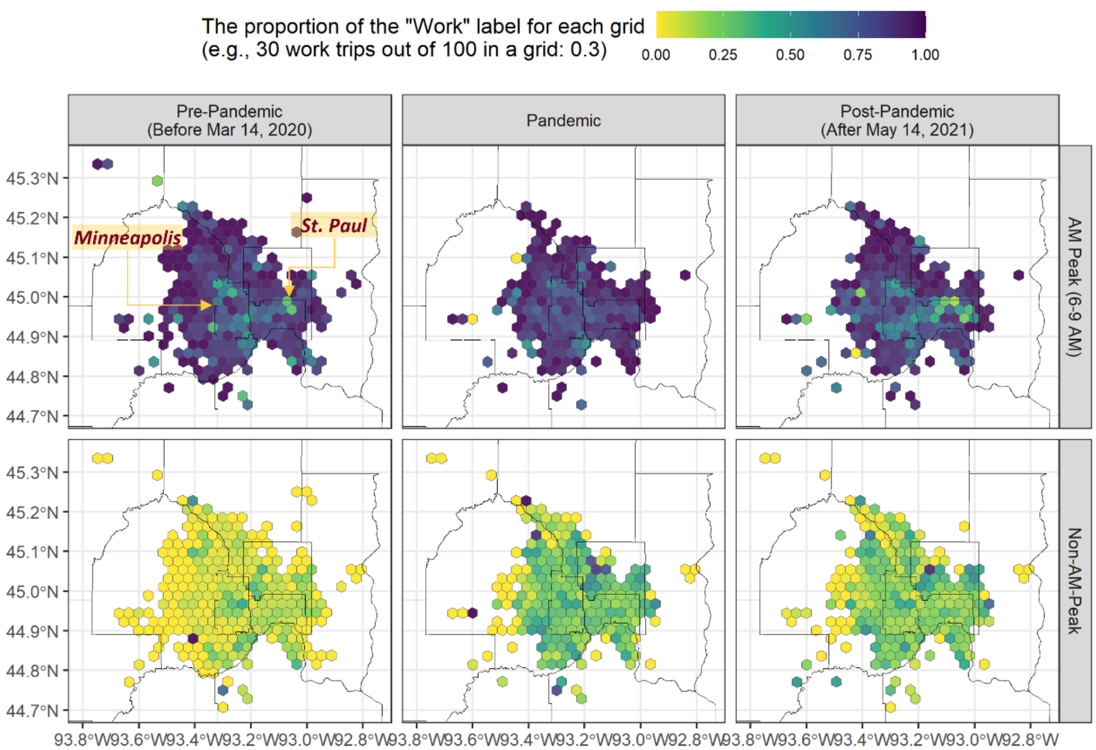


Figure 4.9 Commute trip proportion destined to each grid by time

Figure 4.10 builds upon **Figure 4.9** by focusing on the dominance of "Work" labeled trips based on time of day. Each grid is color-coded to represent the share of AM Peak and Non-AM Peak commute trips, emphasizing the temporal distribution of work-related travels. For example, if a grid has the same number of AM-peak commutes and non-AM-peak commutes identified, it will be white-color coded. Alternatively, if the AM-peak commute dominates the total commute trips in a grid, it will be red. Notably, the post-pandemic panel shows more blue grids than the pre-pandemic period. This again suggests a decrease in the concentration of commutes during traditional peak hours, aligning with the evolving work dynamics observed during and after the pandemic.

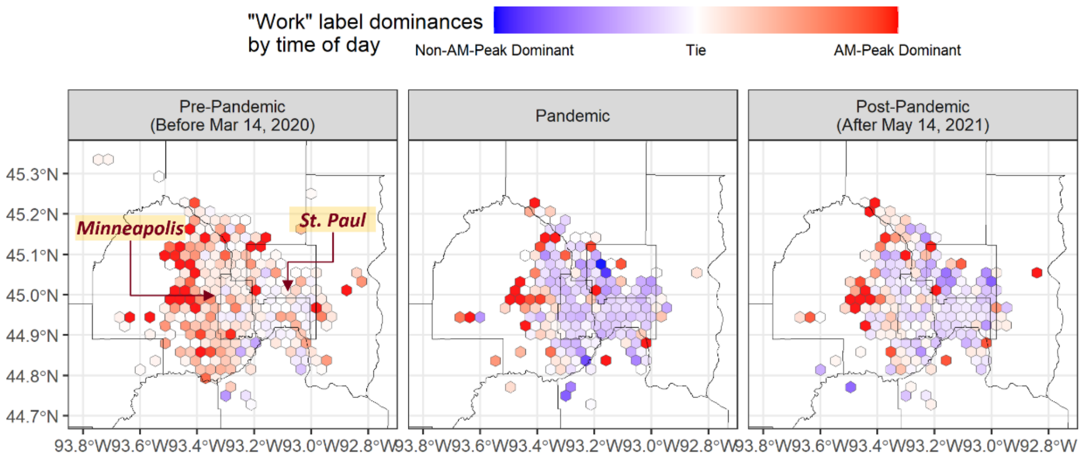


Figure 4.10 Time of day concentration among commute trips destined to each grid

Table 4.3 summarizes the aggregated pre- and post-pandemic proportions depicted in **Figure 4.9** and **Figure 4.10**, categorized by zones (**Figure 2.6**) and Twin Cities region-wide statistics. Zone 16, *Park and Ride Dominant*, is not reported in the table because the number of work trips destined to the zone inferred from the neural network model is small. The findings at both the zonal and regional levels corroborate an increased prevalence of off-peak commute proportions post-pandemic.

Table 4.3 Commute proportion and AM-peak commute proportion by zone

Zone (From Chapter 2)	Commutes (to work)/Total Trips				AM-peak Commutes/Commutes	
	Pre-pandemic		Post-Pandemic		Pre-pandemic	Post-pandemic
	AM Peak	Others	AM Peak	Others		
Downtown Minneapolis	97.7%	31.6%	95.9%	42.9%	77.5%	56.1%
Southwest Minneapolis	73.0%	15.7%	70.7%	24.3%	52.7%	35.0%
Southeast Minneapolis	66.8%	14.4%	66.0%	22.5%	49.8%	37.3%
University of Minnesota	82.0%	32.0%	71.8%	31.5%	48.5%	39.7%
Northeast Minneapolis	76.7%	7.2%	83.4%	16.8%	73.8%	58.7%
North Minneapolis	66.3%	10.6%	71.5%	22.0%	51.5%	34.5%
Downtown St. Paul	94.5%	17.1%	94.9%	34.6%	78.6%	46.4%
Highland-Midway	72.3%	11.8%	60.7%	20.4%	48.8%	33.5%
Southeast Metro	81.1%	9.6%	84.2%	20.8%	55.4%	40.9%
Northeast Metro	76.3%	12.3%	50.9%	20.0%	55.1%	35.5%

North Metro	86.6%	5.0%	89.9%	14.4%	72.6%	53.5%
Northwest Metro	94.9%	5.5%	87.5%	20.5%	86.6%	55.3%
Southwest Metro	88.1%	5.1%	88.9%	23.0%	84.9%	45.7%
MSP Airport & MoA	75.9%	39.1%	80.7%	46.7%	26.3%	27.9%
South Metro	70.7%	10.7%	86.6%	14.0%	52.1%	52.6%
Aggregated Statistics	**89.3%**	**18.4%**	**82.0%**	**27.8%**	**68.0%**	**46.0%**

Figure 4.11 and **Figure 4.12** show the statistics associated with the location of transfers that happened on each grid. The "Peak" here represents the AM and PM peaks collectively (not only AM-Peak) and off-peak the others. **Figure 4.11** represents the time of day dominance of transfer. The pre-pandemic panel indicates that transfers occurring in suburban areas mainly happened at peak times. However, transfers are no longer concentrated during peak times post-pandemic. Finally, **Figure 4.12** illustrates percentage changes in transfer counts by each grid after the pandemic compared to the pre-pandemic level. While most grids show decreased counts of transfers, both peak and off-peak, a few grids mainly include stations of the newly opened BRT routes exhibit increases.

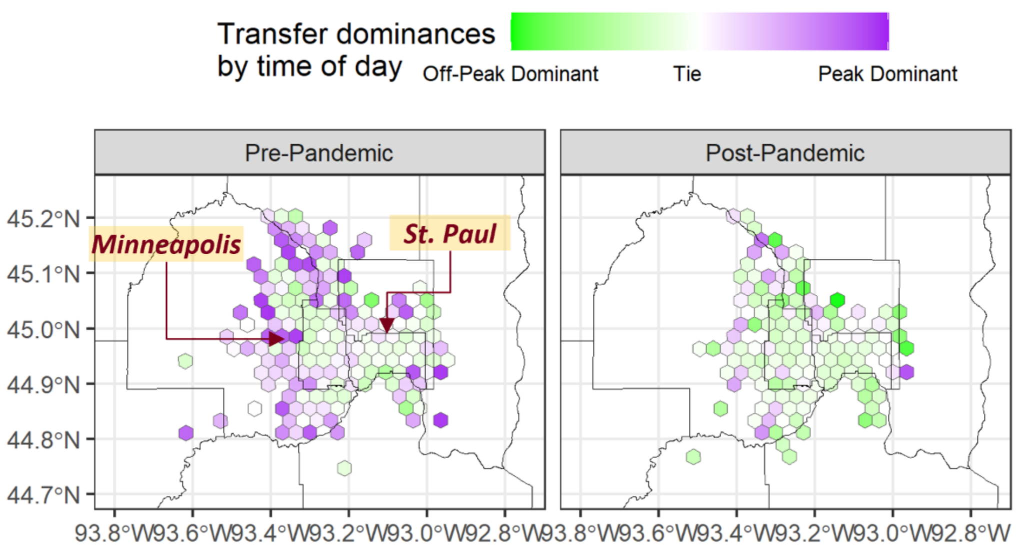


Figure 4.11 A comparison of peak- and off-peak dominance of transfers by periods

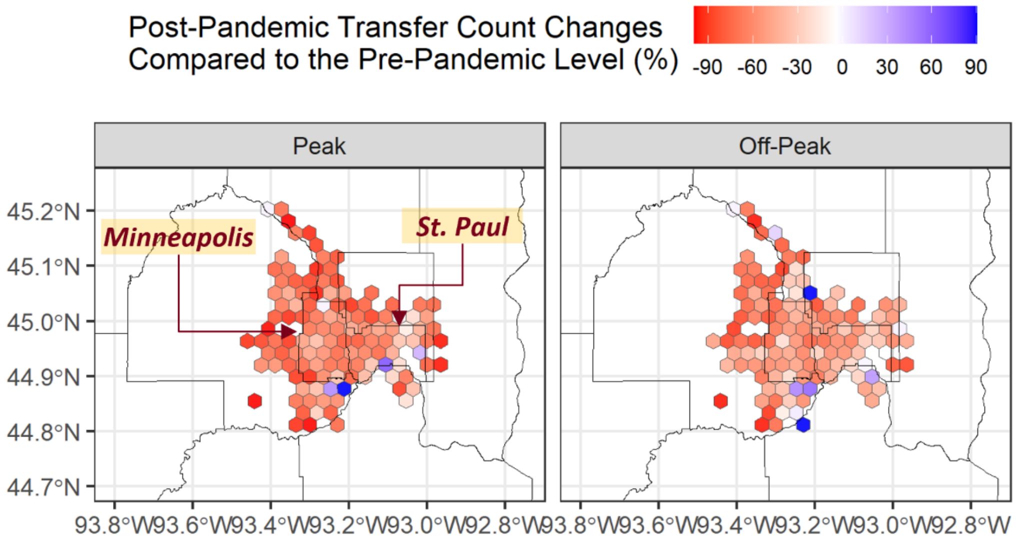


Figure 4.12 A comparison of pre-and post-pandemic transfer count changes by the time of day

It is noteworthy that the ML models were trained on survey data representing trips from 2016. Consequently, there is a possibility that these models may not fully capture behavior changes associated with post-events like the COVID-19 restrictions and alterations in the transit network. Some models heavily reliant on location and time might generate circulatory results (i.e., such models deterministically label trips to downtown during the AM peak as work trips, ignoring the other contexts). Although the selected neural network model demonstrated the least issues regarding the circulatory result, future research could delve into a comprehensive model comparison, potentially incorporating models trained from different domains, such as the Travel Behavior Inventory (TBI) or a more recent version of the onboard survey.

CHAPTER 5: CONCLUSION

In this research project, we explored off-peak commutes and the transfer behavior of transit users. **Chapter 2** encompassed various analyses such as OD place type, travel distance, route type, and HDBSCAN clusters. According to the t-test, peak-time transit trips cover a longer average distance (8.51 mi) compared to off-peak transit travels (5.74 mi). Origin/destination types analysis suggests that during the AM-peak period, home-based trips (92%) and commute trips (72%) are predominant, whereas off-peak times show a higher inclination toward trip purposes such as shopping (62%) and social visits (58%). Spatially, the clustering of trip data indicates that off-peak trips are more dispersed compared to their peak counterparts. Mosaic plots based on the generated zones are then visualized to explore the relationships between various variables and times of departure. Finally, matrix plots further illustrate a concentration of transitway trips and non-home or non-work trip purposes during off-peak hours.

Chapter 3 recovered transitway-including and bus-only path options for eligible survey respondents, constructing path choice sets for analysis. Path choices suggest that transitways play a role in facilitating transfers, with 23% of individuals opting for transitway paths involving transfers even when a direct bus path is available. Logistic regression results reveal that 60% of passengers would prefer their transit path with transitway routes to a bus-only path alternative when travel times are equal. Commuters, especially those using transit during peak times, exhibit lower sensitivity to transportation modes, prioritizing time savings over using transitways.

Chapter 4 integrates the AFC and OBS data, utilizing machine learning techniques to infer trip purposes. Using the inferences, we comprehensively explore transit travel behaviors and long-term transit trip pattern changes in the Twin Cities. In the post-pandemic compared to pre-pandemic, there is a notable increase in the proportion of off-peak commute trips. The proportion rises from 28% to 40% for trips to downtown Minneapolis and St. Paul and from 15% to 23% for trips to other areas. Prior to the pandemic, 68% of transit commutes to workplaces occured during the AM peak, which decreases to 46% post-pandemic. Transfer patterns also exhibit changes, where approximately one-third of transfers are associated with transitways. The total number of transfers decreases after the pandemic, but the proportion of transit trips with transfers increases.

Recognizing the study's limitations is crucial, as they pave the way for future research. First, the path comparison has limitations in accounting for non-walk access modes and the exclusion of short-distance trips. Bias toward long transit trips makes logistic regression results more relevant to such journeys. Second, machine Learning models trained on 2016 survey data have potential gaps in capturing post-event behavioral changes, such as those caused by COVID-19. The issue of circulatory results in models is acknowledged, prompting a call for comprehensive model comparisons trained with more recent years' data.

In conclusion, the identified characteristics and shifts in off-peak commute patterns suggest adaptive transit planning strategies to accommodate evolving behaviors. Valuable insights into transfer dynamics can guide optimizations in routes, schedules, and infrastructure, ultimately enhancing users' overall transit travel experiences. Importantly, establishing a continuous monitoring system becomes paramount for evaluating both short-term responses and the long-term behavior changes of transit users, ensuring that transit services align with the evolving needs of the community. For instance, machine learning models can monitor daily transit trip behavior, identifying any changes. These detected shifts can then become focal points for the subsequent phase of transit planning as a form of a new questionnaire for the next round of the onboard survey. Subsequent analyses of the new survey can delve into the intricate details of transit users' perceptions from the path comparisons and behavioral attributes from the new ML models targeted to the new topic. This iterative process can keep the transit system adaptive and aligned with the dynamic needs of the users.

REFERENCES

Alsger, A., Tavassoli, A., Mesbah, M., Ferreira, L., & Hickman, M. (2018). Public transport trip purpose inference using smart card fare data. *Transportation Research Part C: Emerging Technologies*, *87*, 123–137. https://doi.org/10.1016/j.trc.2017.12.016

Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data*, *10*(1), 1–51. https://doi.org/10.1145/2733381

Ceder, A. (2015). *Public transit planning and operation: Modeling, practice and behavior*. Boca Raton, FL CRC Press. https://books.google.com/books?id=b84dCgAAQBAJ

Chapleau, R., Gaudette, P., & Spurr, T. (2018). Strict and deep comparison of revealed transit trip structure between computer-assisted telephone interview household travel survey and smart cards. *Transportation Research Record: Journal of the Transportation Research Board*, *2672*(42), 13–22. https://doi.org/10.1177/0361198118758297

de Dios Ortúzar, J., & Willumsen, L. G. (2011). *Modelling transport*. Hoboken, NJ: John Wiley & Sons.

ESRI. (2021). *How build balanced zones work*. Retrieved from https://pro.arcgis.com/en/pro-app/3.1/tool-reference/spatial-statistics/learnmore-buildbalancedzones.htm

Fan, W., & Machemehl, R. B. (2009). Do transit users just wait for buses or wait with strategies? Some numerical results that transit planners should see. *Transportation Research Record*, *2111*(1), 169–176. https://doi.org/10.3141/2111-19

Guo, Z., & Wilson, N. H. M. (2011). Assessing the cost of transfer inconvenience in public transport systems: A case study of the London Underground. *Transportation Research. Part A, Policy and Practice*, *45*(2), 91–104. https://doi.org/10.1016/j.tra.2010.11.002

Hensher, D. A. (2008). Assessing systematic sources of variation in public transport elasticities: Some comparative warnings. *Transportation Research. Part A, Policy and Practice*, *42*(7), 1031–1042. https://doi.org/10.1016/j.tra.2008.02.002

Hörcher, D., Graham, D. J., & Anderson, R. J. (2017). Crowding cost estimation with large scale smart card and vehicle location data. *Transportation Research Part B: Methodological*, *95*, 105–125. https://doi.org/10.1016/j.trb.2016.10.015

Khani, A., Hickman, M., & Noh, H. (2015). Trip-based path algorithms using the transit network hierarchy. *Networks and Spatial Economics*, *15*(3), 635–653. https://doi.org/10.1007/s11067-014-9249-3

Koushik, A. N., Manoj, M., & Nezamuddin, N. (2020). Machine learning applications in activity-travel behavior research: A review. *Transport Reviews*, *40*(3), 288–311. https://doi.org/10.1080/01441647.2019.1704307

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

Lee, S. G., & Hickman, M. (2014). Trip purpose inference using automated fare collection data. *Public Transport*, *6*(1–2), 1–20. https://doi.org/10.1007/S12469-013-0077-5/FIGURES/6

Metropolitan Council. (2015). *2040 Transportation policy plan*. St. Paul, MN: Metropolitan Council.

Metropolitan Council. (2016). *Regional transitway guidelines*. St. Paul, MN: Metropolitan Council.

Meyer, D., Zeileis, A., & Hornik, K. (2006). The strucplot framework: Visualizing multi-way contingency tables with vcd. *Journal of Statistical Software*, *17*(1), 1–48. https://doi.org/10.18637/jss.v017.i03

Parker, M. E. G., Li, M., Bouzaghrane, M. A., Obeid, H., Hayes, D., Frick, K. T., … & Chatman, D. G. (2021). Public transit use in the United States in the era of COVID-19: Transit riders' travel behavior in the COVID-19 impact and recovery period. *Transport Policy*, *111*, 53–62. https://doi.org/10.1016/j.tranpol.2021.07.005

Raveau, S., Guo, Z., Muñoz, J. C., & Wilson, N. H. M. (2014). A behavioral comparison of route choice on metro networks: Time, transfers, crowding, topology and socio-demographics. *Transportation Research. Part A, Policy and Practice*, *66*(1), 185–195. https://doi.org/10.1016/j.tra.2014.05.010

Tomhave, B. J., & Khani, A. (2022). Refined choice set generation and the investigation of multi-criteria transit route choice behavior. *Transportation Research. Part A, Policy and Practice*, *155*, 484–500. https://doi.org/10.1016/j.tra.2021.11.005

Train, Kenneth. (2009). *Discrete choice methods with simulation*. Cambridge, UK: Cambridge University Press.

Vuchic, V. R. (2005). *Urban transit: Operations, planning, and economics*. Hoboken, NJ: John Wiley & Sons.

Wang, D., Tayarani, M., He, B. Y., Gao, J., Chow, J. Y. J., Gao, H. O., & Ozbay, K. (2021). Mobility in post-pandemic economic reopening under social distancing guidelines: Congestion, emissions, and contact exposure in public transit. *Transportation Research Part A: Policy and Practice*, *153*, 151–170. https://doi.org/10.1016/J.TRA.2021.09.005

Webb, A., Kumar, P., & Khani, A. (2020). Estimation of passenger waiting time using automatically collected transit data. *Public Transport*, *12*(2), 299–311. https://doi.org/10.1007/s12469-020-00229-x

Zhao, D., Wang, W., Woodburn, A., & Ryerson, M. S. (2017). Isolating high-priority metro and feeder bus transfers using smart card data. *Transportation*, *44*(6), 1535–1554. https://doi.org/10.1007/s11116-016-9713-7

Zhu, W., Fan, W., Wahaballa, A. M., & Wei, J. (2020). Calibrating travel time thresholds with cluster analysis and AFC data for passenger reasonable route generation on an urban rail transit network. Transportation, 47(6), 3069–3090. https://doi.org/10.1007/s11116-019-10040-8

APPENDIX A

MOSAIC PLOTS

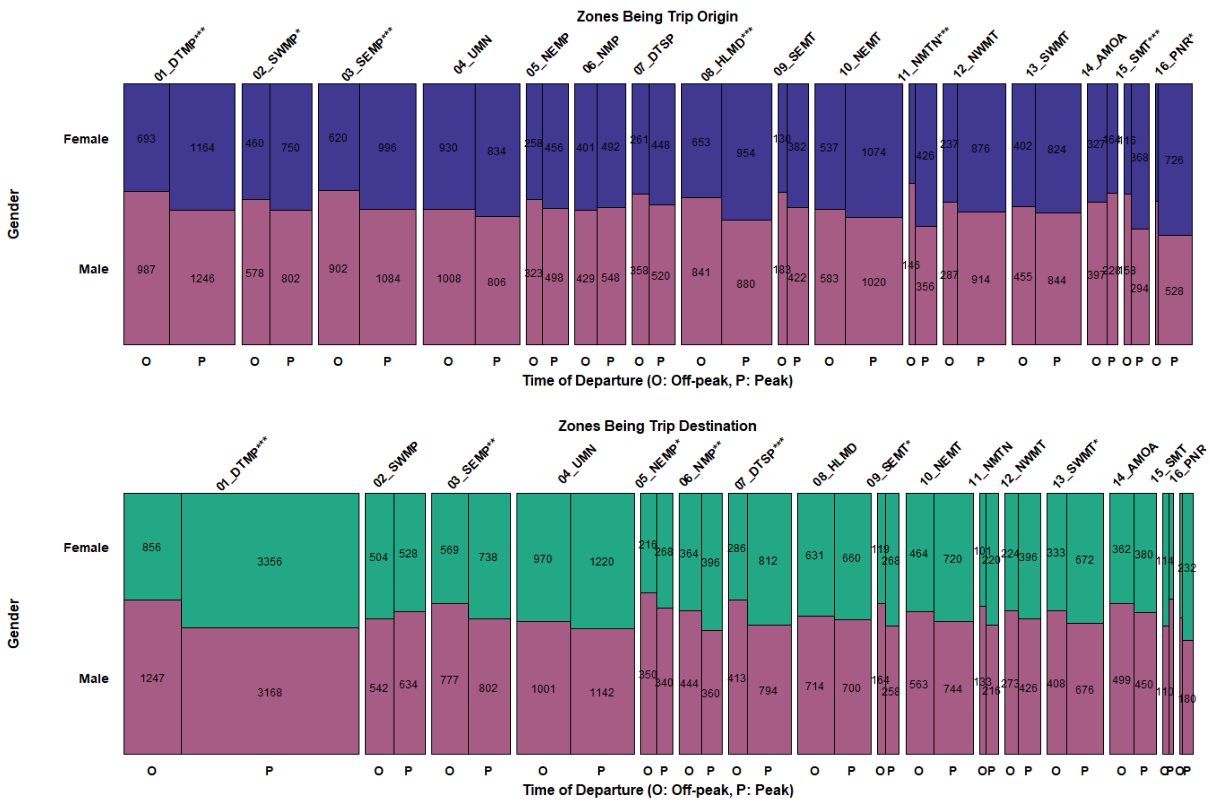
This appendix provides additional mosaic plots summarized in the table below.

Inspected Variable	Variable Category (count)	When Zones are Being	Zone Numbers whose χ^2 test showed statistically significant results (count)
Gender	•Female (17,088) •Male (18,673)	Origin	1, 2, 3, 8, 11, 15, 16 (7 zones)
		Destination	1, 3, 5, 6, 7, 9, 13 (7 zones)
Age Group	•Age over 44 (10,889) •Age 25-44 (15,588) •Age under 25 (9,284)	Origin	1, 2, 4, 5, 7, 8, 10 – 16 (13 zones)
		Destination	1, 2, 4, 7, 11, 12, 13, 16 (8 zones)
Access Mode	•Motorized (5,735) •Walking/Cycling (30,026)	Origin	5, 6, 8–13, 15, 16 (10 zones)
		Destination	1, 3, 4, 5, 7, 12, 13, 14 (8 zones)
Egress Mode	•Motorized (2,090) •Walking/Cycling (33,671)	Origin	1, 6, 7 (3 zones)
		Destination	1, 3, 7, 10, 11, 13 (6 zones)
Car Availability	•Available (13,747) •Not Available (22,014)	Origin	All zones except Zone 14 (15 zones)
		Destination	All zones except Zones 2 and 15 (14 zones)
Payment Method	•Cash (7,935) •Freeride or Pass used (14,883) •Go-to Card or Mobile (12,943)	Origin	1–16 (all 16 zones)
		Destination	1–16 (all 16 zones)
Transfers	•Not transferred (21,067) •Transferred (14,694)	Origin	1, 2, 4, 8–15 (11 zones)
		Destination	1, 2, 3, 4, 13, 15 (6 zones)
Household	•Over $60K (16,814) •$15K~$60K (14,381)	Origin	All zones except Zone 14 (15 zones)

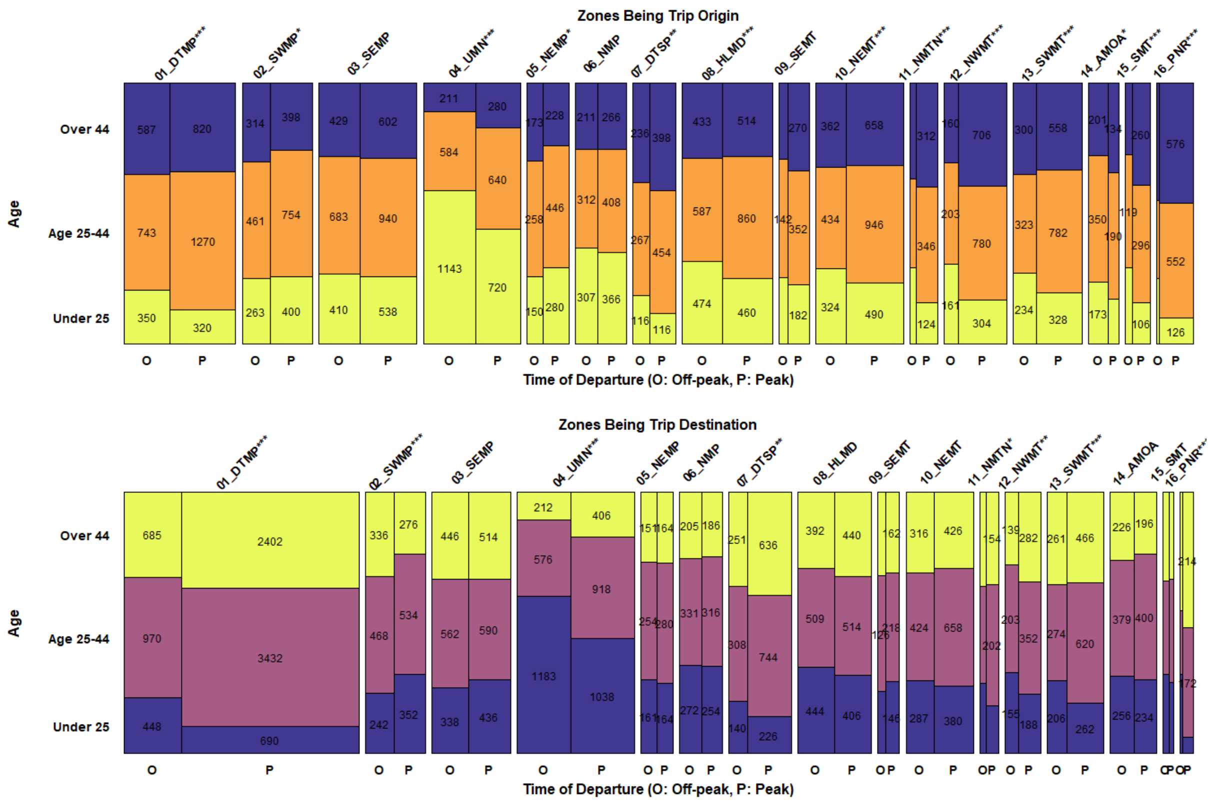
Income (Annual) (**Figure 2.10**)	•Under $15K (4,566)	Destination	All zones except Zone 15 (15 zones)
Employment/ student status	•Student (10,364) •Employed non-student (22,402) •None of the above (2,995)	Origin	All zones except Zone 16 (15 zones)
		Destination	All zones except Zone 15 (15 zones)
Disability	•With a disability (3,837) •Without disability (31,924)	Origin	All zones except Zones 4 and 14 (14 zones)
		Destination	1, 2, 3, 6–13 (11 zones)
Race	•White identifying (22,079) •Non-white (13,682)	Origin	All zones except Zones 6 and 14 (14 zones)
		Destination	1, 3, 4, 7, 8, 10, 11, 16 (8 zones)
Trip Distance	•Over 10mi (8,156) •Under 10mi (27,605)	Origin	1, 4, 7, 9–13, 15 (9 zones)
		Destination	1, 3, 4, 7, 8, 11, 13, 14 (8 zones)
Home Language	•English (30,336) •Others (5,425)	Origin	5, 6, 8, 11, 15, 16 (6 zones)
		Destination	1, 3, 4, 8, 16 (5 zones)
Transitway (**Figure 2.9**)	•Including Transitway (11,863) •Not Including (23,898)	Origin	1, 4, 7, 8, 10, 12, 15, 16 (8 zones)
		Destination	1, 2, 4–10, 12, 13, 16 (12 zones)
Household Member Count	•One (7,085) •Two (10,180) •More than two (18,496)	Origin	1–5, 8, 10–13, 15, 16 (12 zones)
		Destination	1, 2, 3, 4, 7, 8, 10, 12, 13, 14 (10 zones)
Fare Subsidy	•Full (3,280) •Partial (10,146) •No Subsidy (22,335)	Origin	1–16 (all 16 zones)
		Destination	All zones except Zones 14 and 15 (14 zones)
	•Home (22,359) •Work (6,467)	Origin	All zones except Zone 15 (15 zones)

Origin Place Type	•Others (6,935)	Destination	1–16 (all 16 zones)
Destination Place Type	•Home (9,793) •Work (15,004) •Others (10,964)	Origin	1–16 (all 16 zones)
		Destination	All zones except Zones 9 and 16 (14 zones)

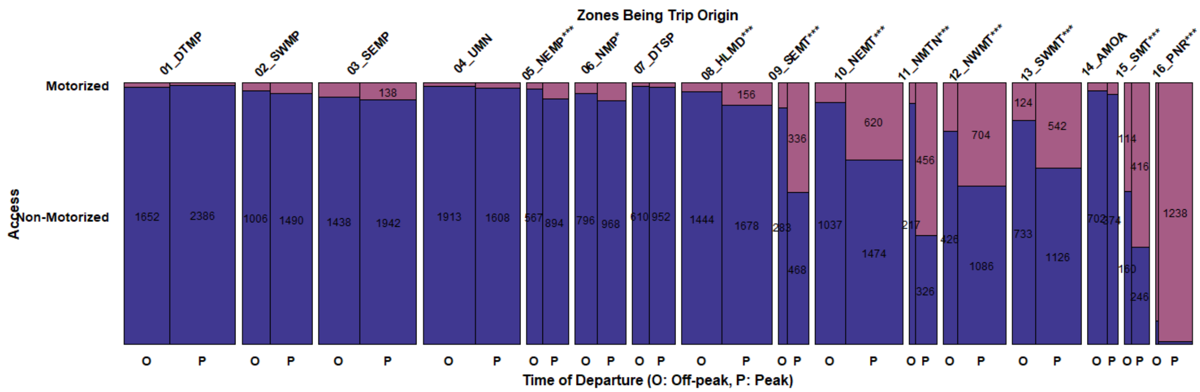
* Some zones are excluded in the test due to the small sample size less than 5/cell (marked "?" in plots)

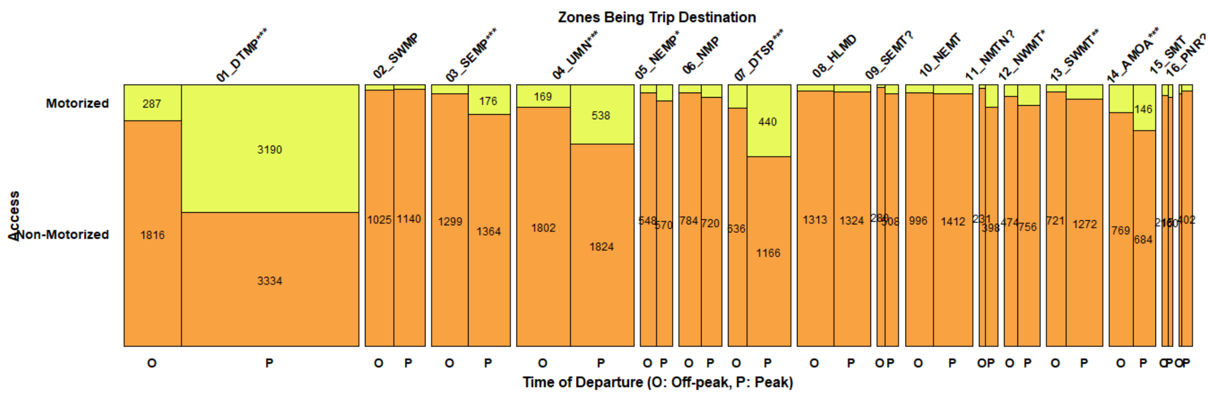


Mosaic plots: gender

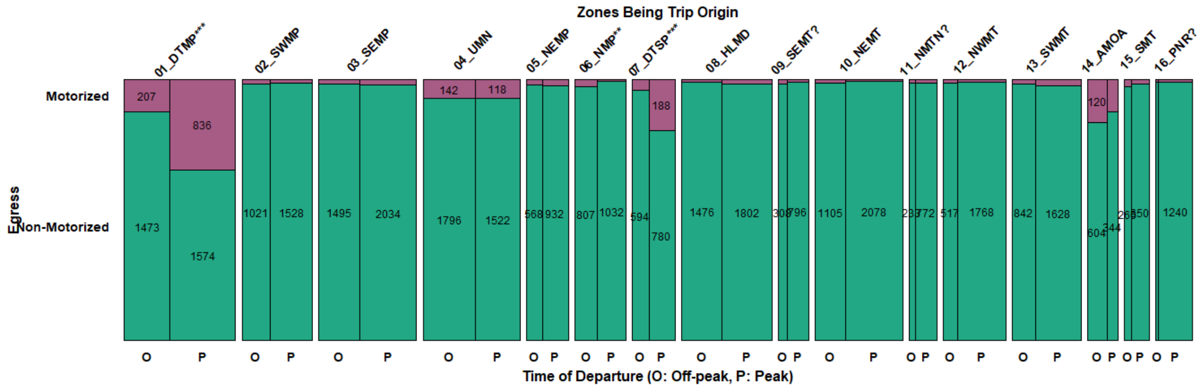


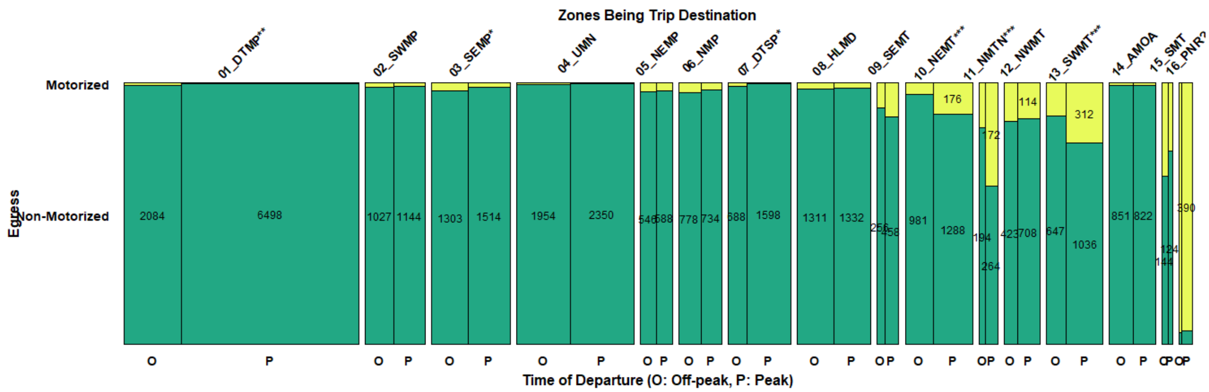
Mosaic plots: age



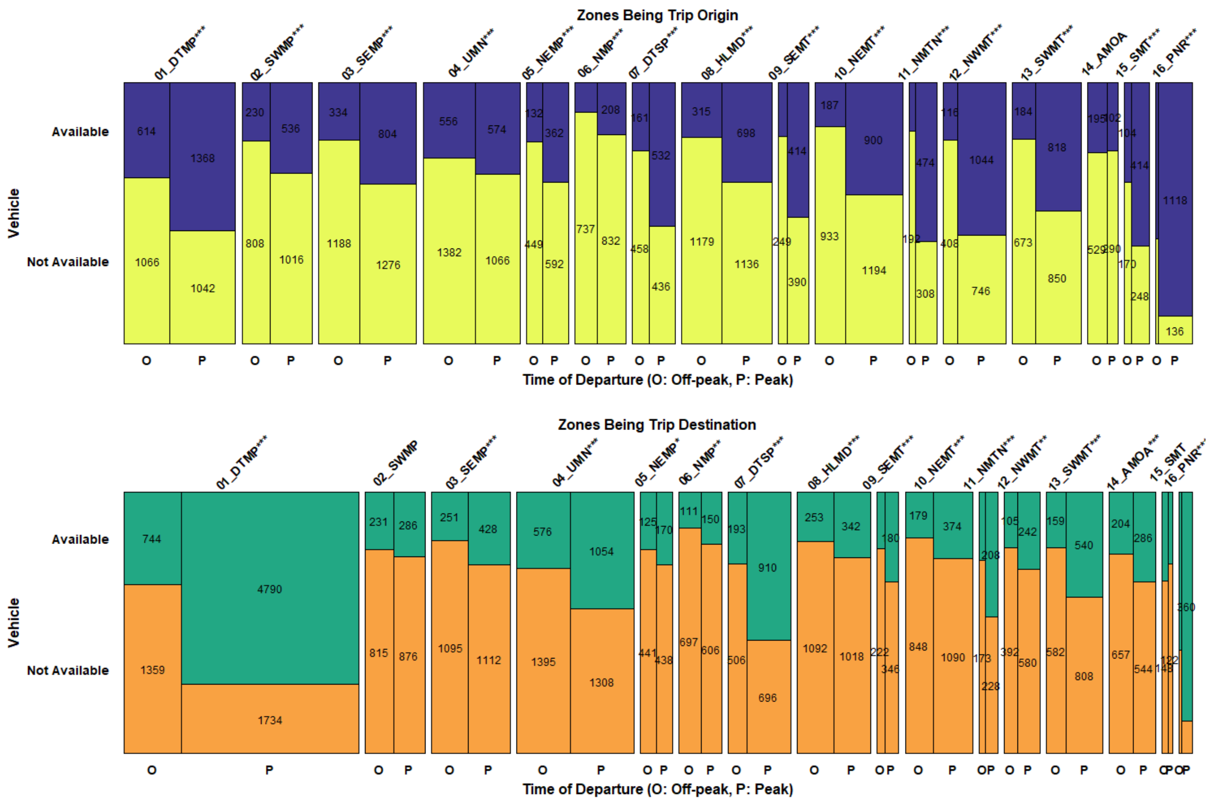


Mosaic plots: access mode

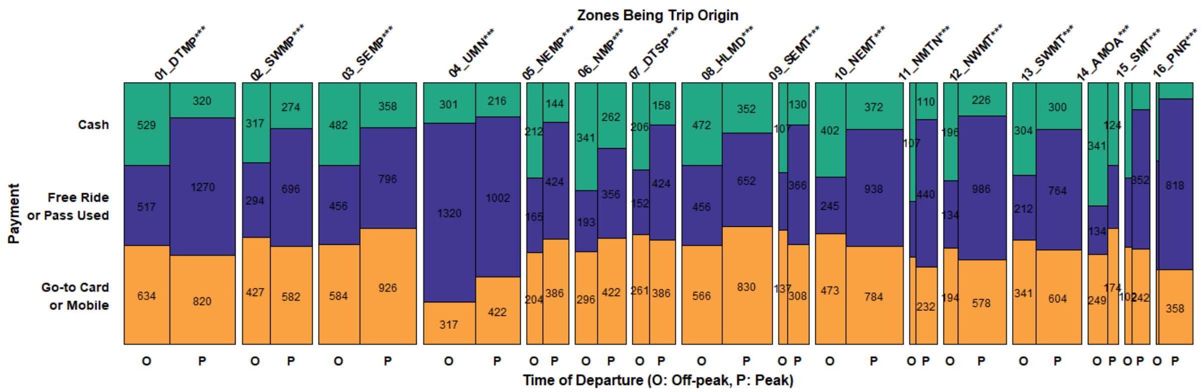


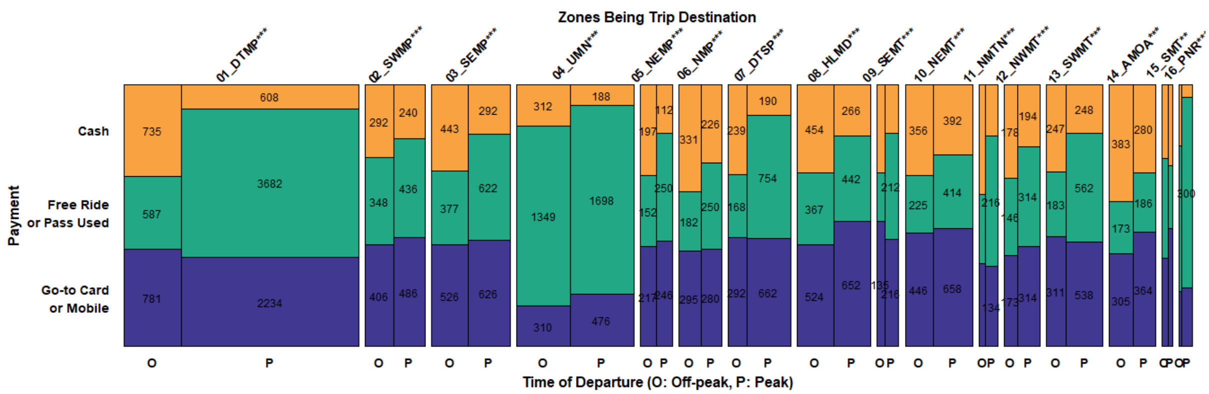


Mosaic plots: egress mode

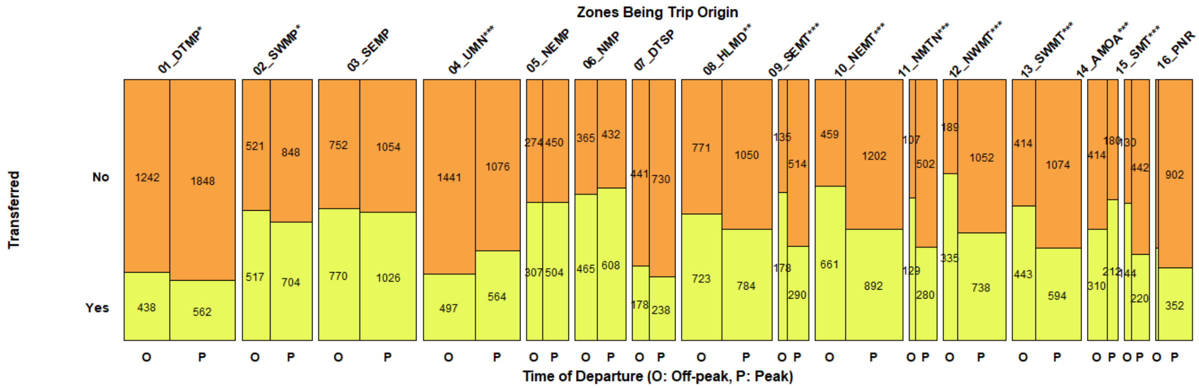


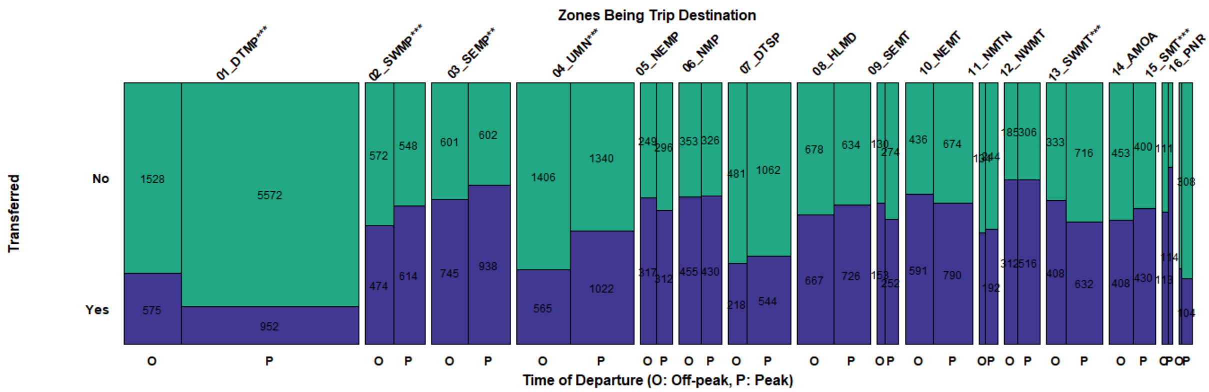
Mosaic plots: household vehicle availability



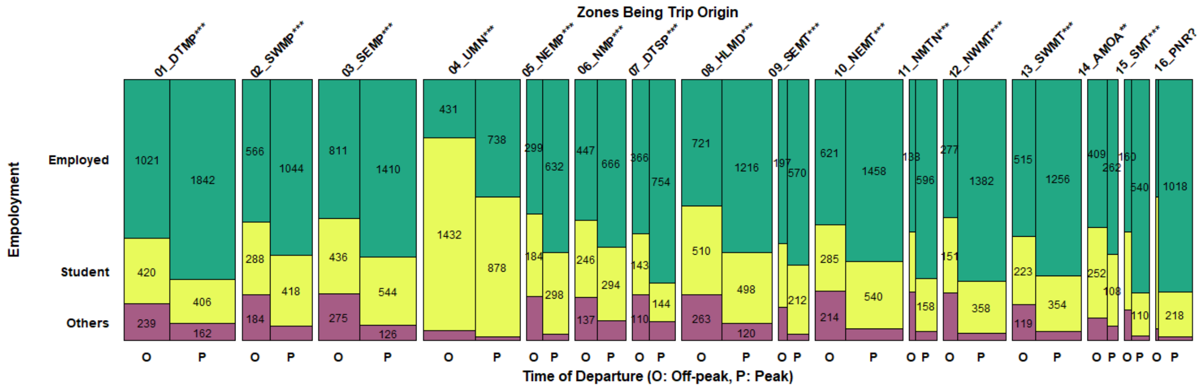


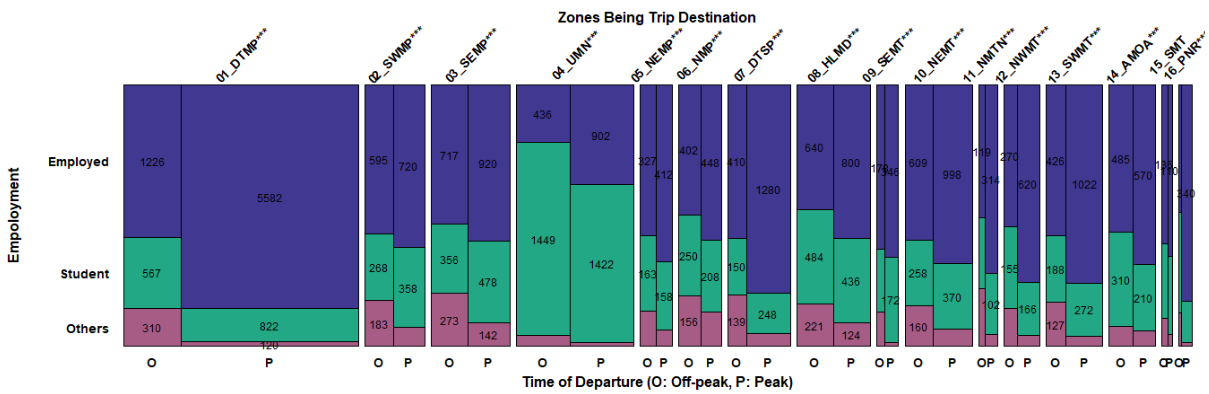
Mosaic plots: fare payment method



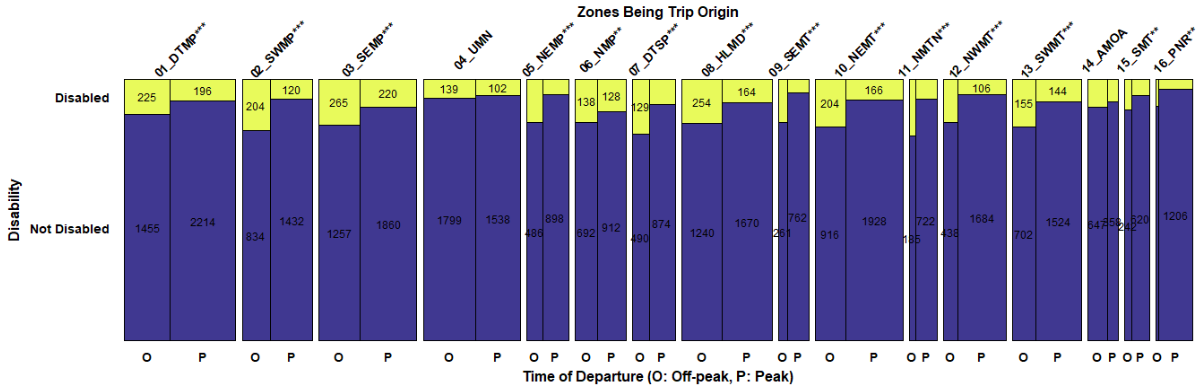


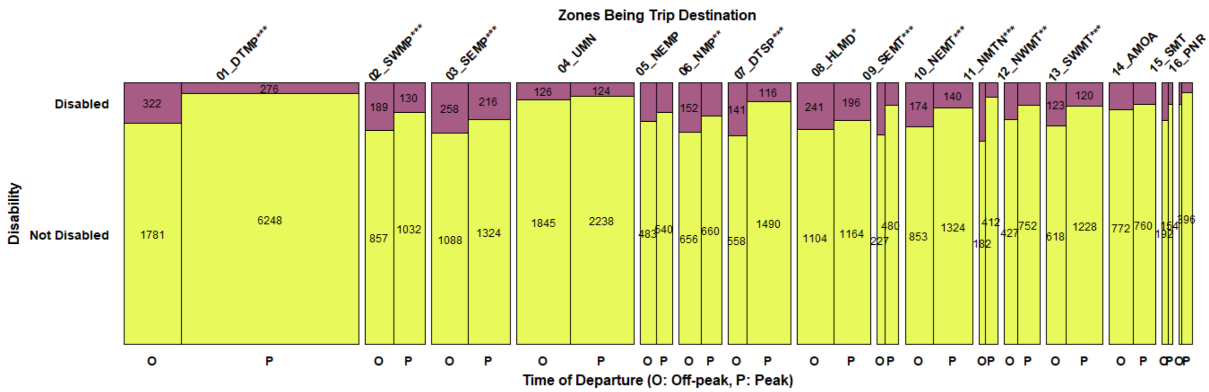
Mosaic plots: made transit transfer(s) during the trip



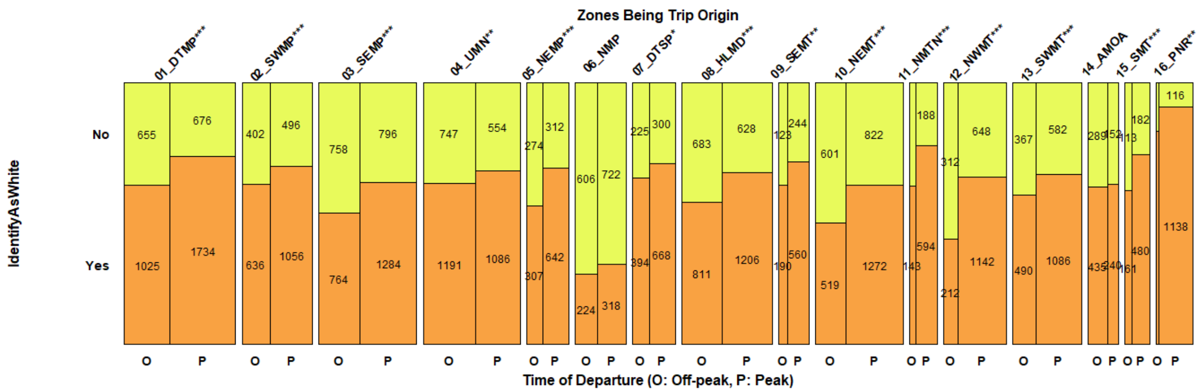


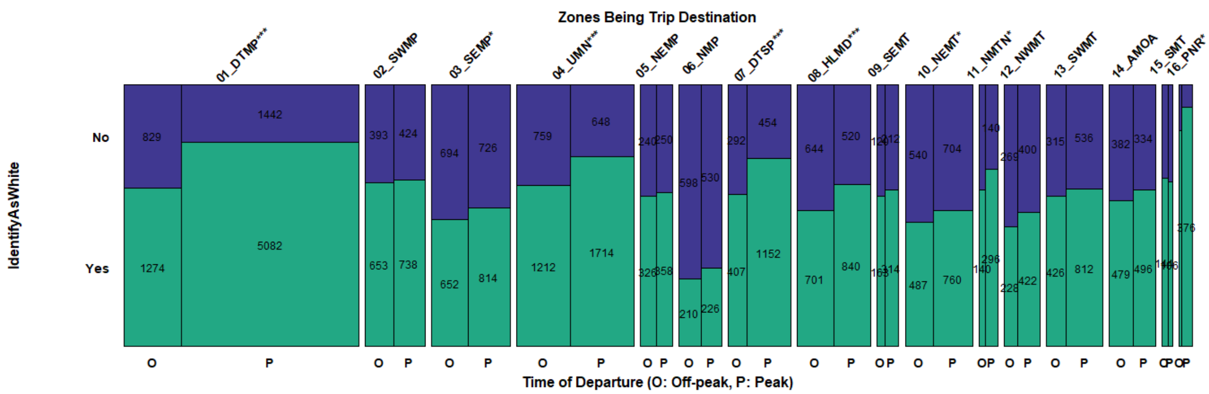
Mosaic plots: employment/student status



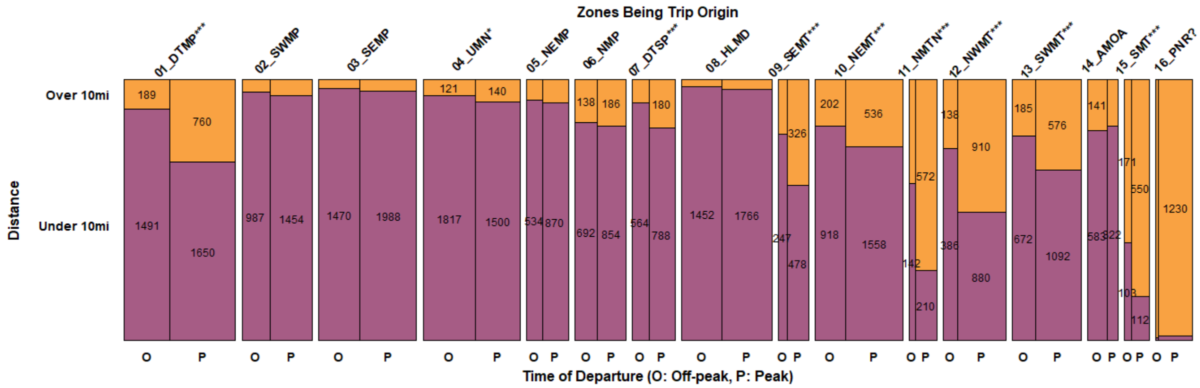


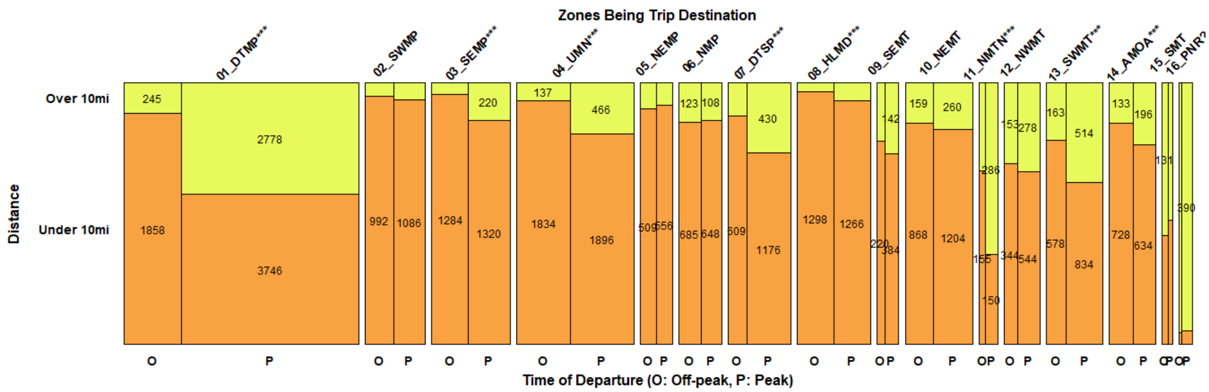
Mosaic plots: disability



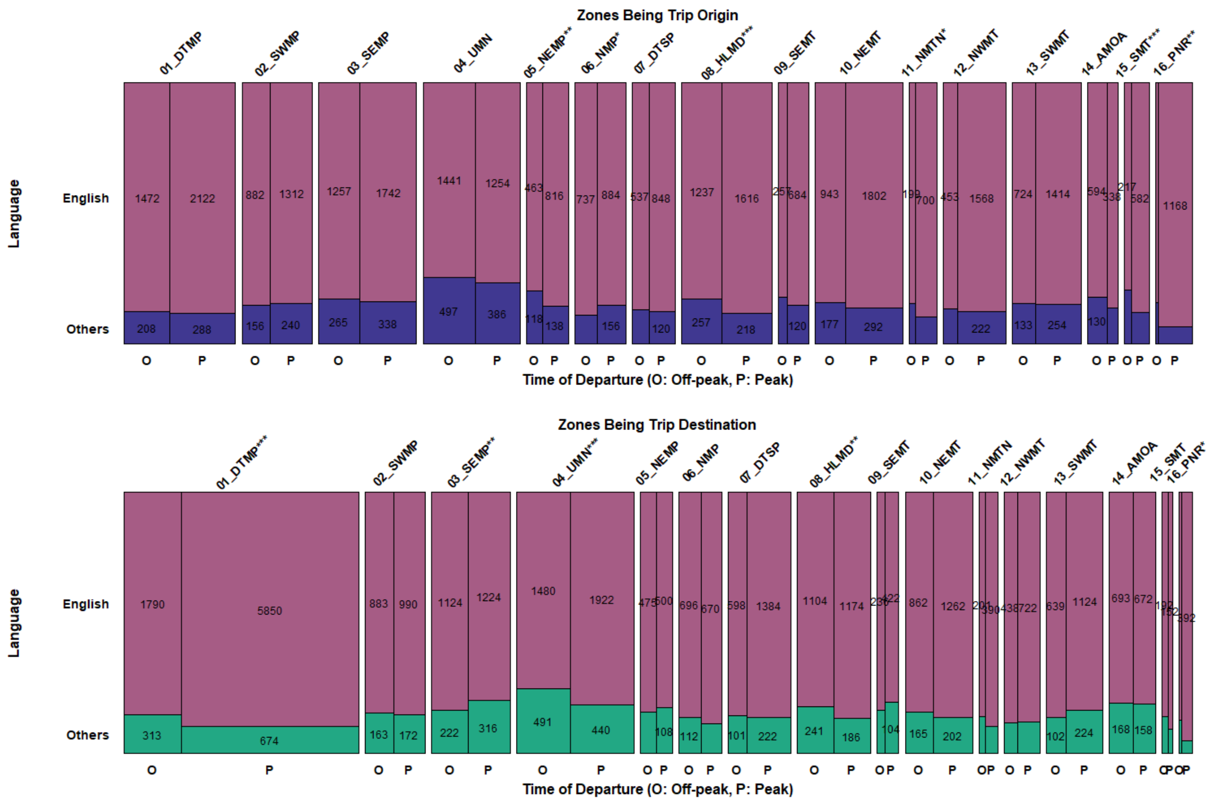


Mosaic plots: race

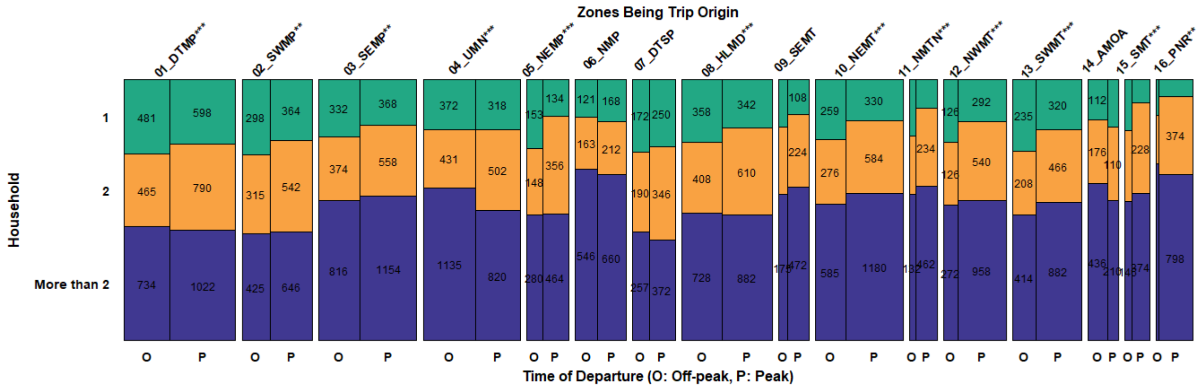


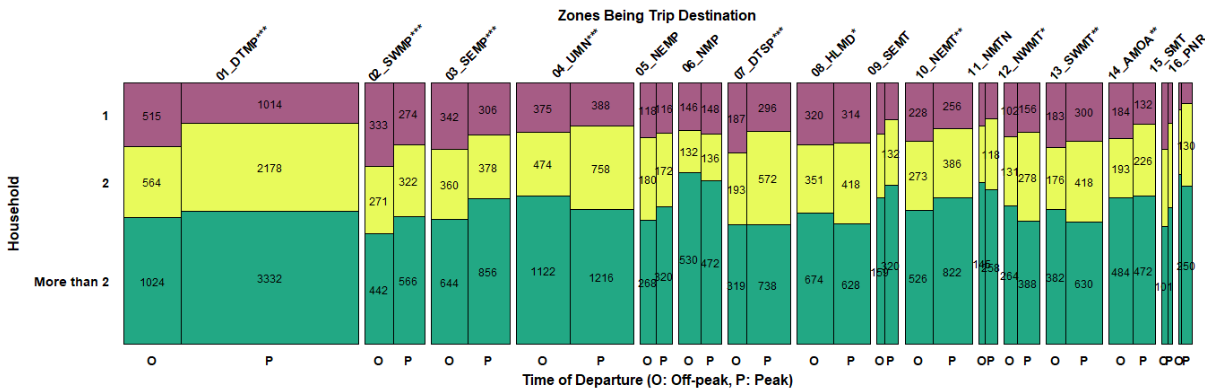


Mosaic plots: trip distance (Euclidean)

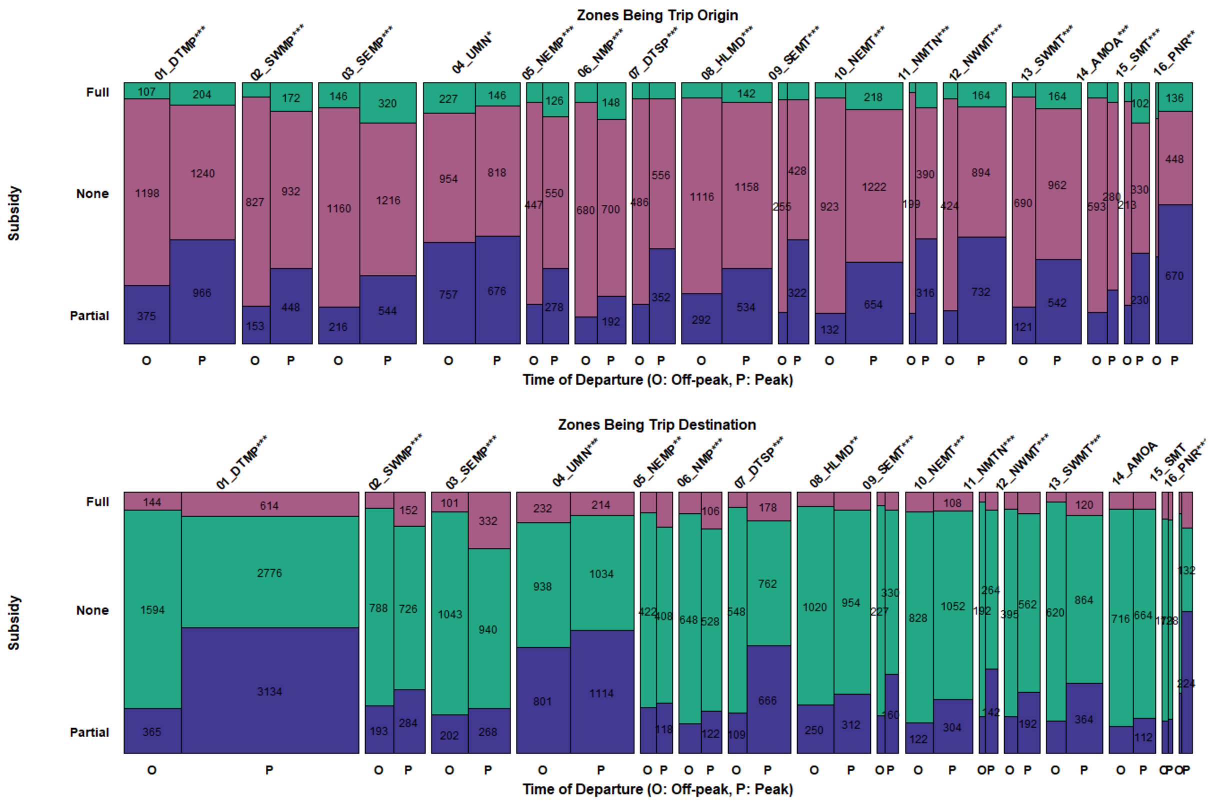


Mosaic plots: home language

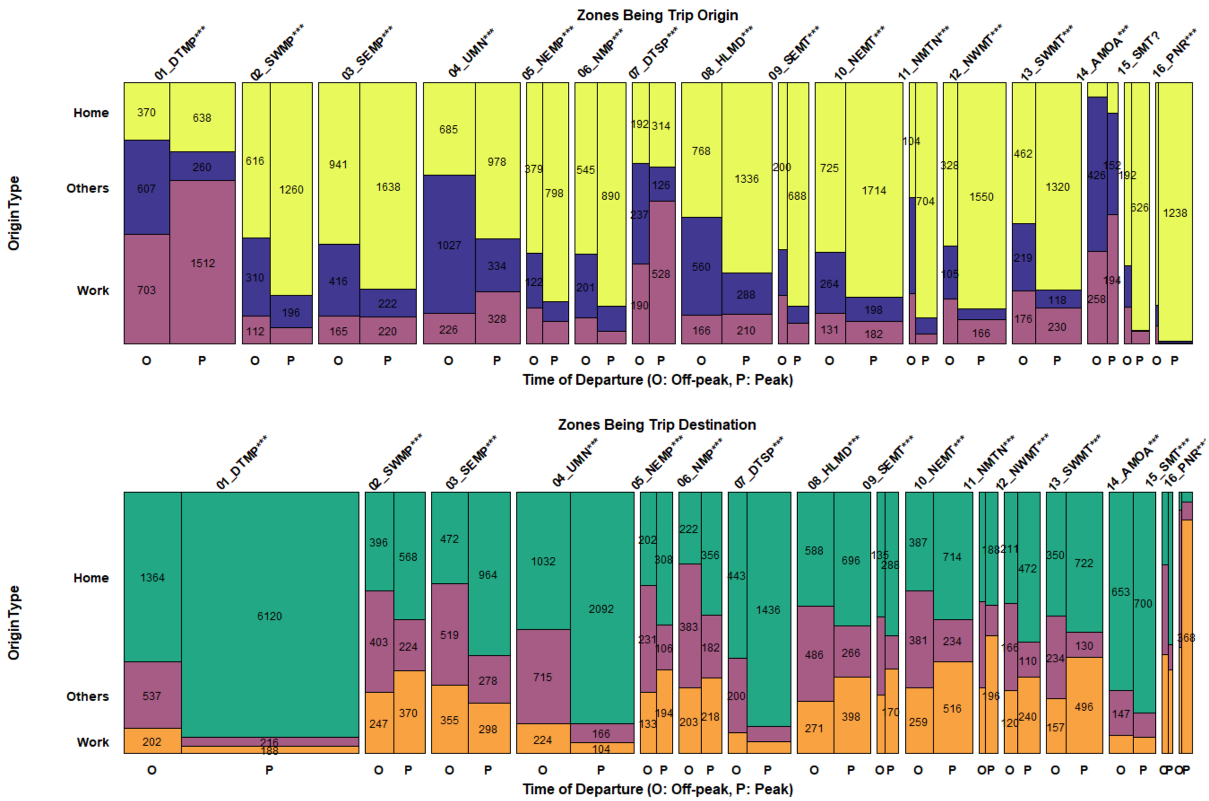




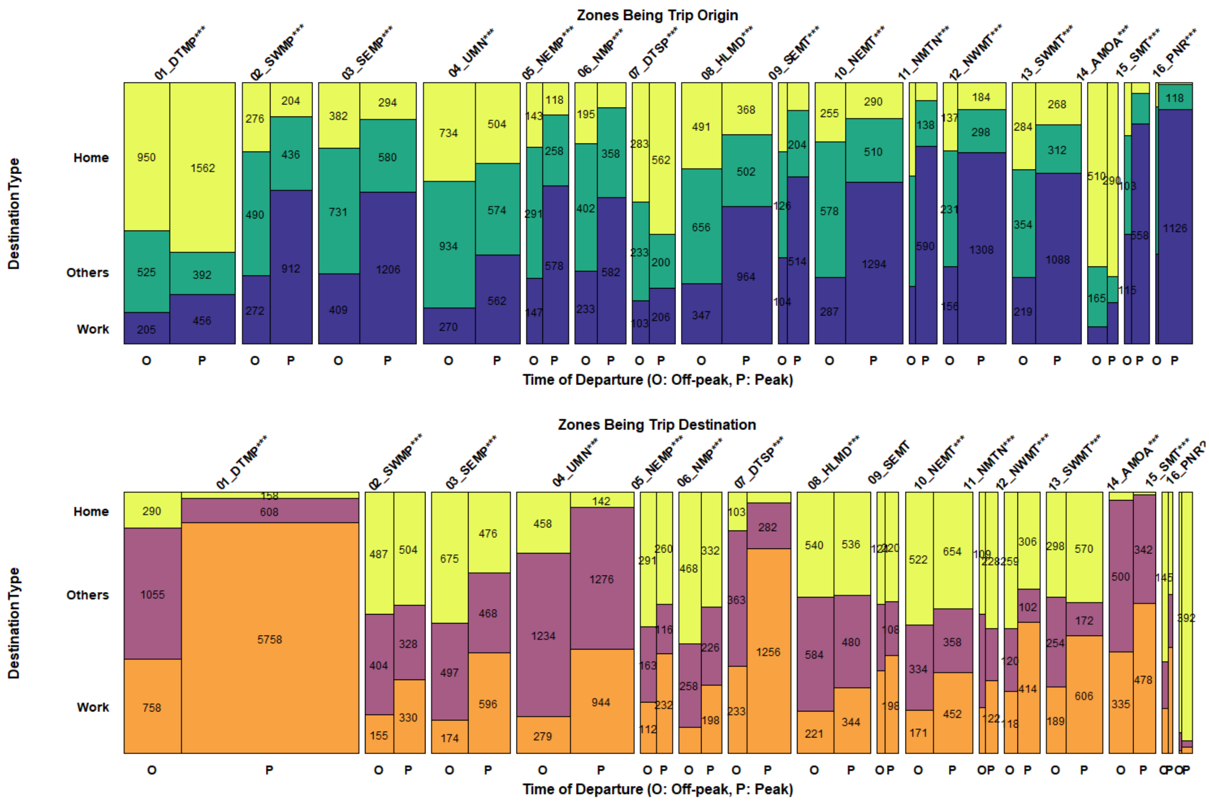
Mosaic plots: the number of household members



Mosaic plots: fare subsidy



Mosaic plots: origin place type



Mosaic plots: destination place type