CENTER FOR CONNECTED
AND AUTOMATED
TRANSPORTATION

Report No. CCAT-NU-2026-1 January 2026

Project Start Date: 7/1/2024

Project End Date: 1/31/2026

Learning Transportation Insecurity from Location Intelligence Data

Yu (Marco) Nie

Ying Chen

Alexandra K. Murphy

Tianxing Dai

Xiaoyu Yan

Peeter Kivestu

Northwestern University
University of Michigan

CCAT | CENTER FOR CONNECTED
AND AUTOMATED
TRANSPORTATION

DISCLAIMER

Funding for this research was provided by the Center for Connected and Automated Transportation under Grant No. 69A3552348305 of the U.S. Department of Transportation, Office of the Assistant Secretary for Research and Technology (OST-R), University Transportation Centers Program. The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Suggested APA Format Citation:

Nie, Y., Chen, Y., Murphy, A., Dai, T., Yan, X., Kivestu, P. (2026) Learning Transportation Insecurity from Location Intelligence Data. CCAT Report *CCAT-NU-2026-1*, The Center of Connected and Automated Transportation, Northwestern University, Evanston, IL.

Contacts

For more information:

Yu (Marco) Nie
Northwestern University
2145 Sheridan Road
Evanston, IL 60201
y-nie@northwestern.edu
(847) 467-0502

CCAT
University of Michigan Transportation Research Institute
2901 Baxter Road
Ann Arbor, MI 48152
uumtri-ccat@umich.edu
(734) 763-2498

M | UMTRI   The University of Akron   ILLINOIS CENTER FOR TRANSPORTATION   PURDUE UNIVERSITY   Washtenaw Community College   Central State University 1887

CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

Technical Report Documentation Page

1. Report No. CCAT-NU-2026-1

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
Learning Transportation Insecurity from Location Intelligence Data

5. Report Date
January 2026

6. Performing Organization Code N/A

7. Author(s)
Yu (Marco) Nie (http://orcid.org/0000-0003-2083-470X),
Ying Chen (https://orcid.org/0000-0001-8669-9366),
Alexandra K. Murphy (https://orcid.org/0000-0002-5786-3072),
Tianxing Dai (http://orcid.org/0000-0003-1984-1857),
Xiaoyu Yan (http://orcid.org/0000-0002-4542-9308),
Peeter Kivestu (https://orcid.org/0009-0007-9320-1168)

8. Performing Organization Report No.
N/A

9. Performing Organization Name and Address

Northwestern University
2145 Sheridan Road
Evanston, IL, 60201

University of Michigan
735 South State Street,
Ann Arbor, MI 48109

10. Work Unit No.

11. Contract or Grant No.
Contract No. 69A3552348305

12. Sponsoring Agency Name and Address
Center for Connected and Automated Transportation
University of Michigan Transportation Research Institute
2901 Baxter Road
Ann Arbor, MI 48109

13. Type of Report and Period Covered
Final Report (July 2024 – January 2026)

14. Sponsoring Agency Code
OST-R

15. Supplementary Notes
Conducted under the U.S. DOT Office of the Assistant Secretary for Research and Technology's (OST-R) University Transportation Centers (UTC) program.

16. Abstract

Transportation insecurity (TI)— a condition in which one is unable to regularly move from place to place in a safe or timely manner because one lacks the material, economic, or social resources necessary for transportation—remains a persistent barrier to economic opportunity and quality of life in the United States. While the Transportation Security Index (TSI) provides a validated measure of TI, its reliance on stand-alone surveys limits its scalability, temporal coverage, and geographic comparability. This report proposes a scalable alternative for measuring TI by leveraging large-scale location intelligence data and transfer learning methods. Building on evidence that mode use patterns are strongly associated with TI status, we develop a framework that infers TI from passively collected mobility trajectories. The approach first employs a semi-supervised machine learning pipeline to identify driving, transit, and active travel modes from sparse and unlabeled mobile phone data, integrating trip-level features with mode-specific travel characteristics from the Google Maps API. We then apply a TI classification model trained on survey data from the Detroit Metro Area Communities Study to location intelligence data from Chicago. Despite relying on simplified mode categories and census tract-level socio-demographic inputs, the model achieves approximately 70% accuracy and recall. While cross-region transfer introduces some bias, the resulting spatial patterns of inferred transportation insecurity align with well-documented areas of mobility disadvantage. These findings demonstrate the feasibility and promise of using location intelligence data to measure transportation insecurity at scale, enabling more systematic monitoring and evaluation of transportation equity outcomes

17. Key Words
Transportation insecurity, mobile phone data, transfer learning, classification, mode inference

18. Distribution Statement
No restrictions.

19. Security Classif. (of this report)
Unclassified

20. Security Classif. (of this page)
Unclassified

21. No. of Pages
43

22. Price

Contents

List of Figures

List of Tables

1 Introduction

Millions of Americans experience significant mobility constraints that limit their ability to access employment, education, healthcare, and other essential services. These challenges reduce economic productivity and quality of life and pose persistent concerns for transportation planning and public policy. Many of these constraints reflect long-standing patterns in transportation investment and land-use decisions, which have resulted in uneven levels of service quality and accessibility across population groups and regions. Addressing these mobility gaps requires improved data, analytical methods, and measurement tools that al-low policymakers to systematically assess transportation access, reliability, and safety. Such tools are essential for the federal government to design, implement, and evaluate transportation policies and programs concerning adequate and cost-effective service provision in the transportation sector.

One such tool developed in recent years is the Transportation Security Index (TSI), a validated measure of transportation insecurity (TI), defined as "a condition in which one is unable to regularly move from place to place in a safe or timely manner because one lacks the material, economic, or social resources necessary for transportation" (Gould-Werth et al., 2018; Murphy et al., 2021). Modeled after the Food Security Index, the TSI measures individuals' experiences with TI by asking survey respondents how frequently they encounter a set of transportation-related difficulties identified through prior qualitative research. The index classifies individuals into multiple categories of transportation security (e.g., secure, low insecurity, high insecurity) (McDonald-Lopez et al., 2023). It has been used to generate the first national estimates of transportation insecurity in the United States and to document systematic variation by income, geography, and demographic characteristics (Murphy et al., 2024). These findings demonstrate the TSI's usefulness as an empirical tool for diagnosing transportation access challenges and evaluating policy interventions.

While the TSI has proven useful in helping planners, researchers, and policymakers understand how transportation disadvantages vary across demographic groups and geographic areas, it has not yet been incorporated into any recurring, nationally representative surveys with large sample sizes. Instead, applications of the TSI to date have relied on stand-alone data collection efforts led by individual researchers or Metropolitan Planning Organizations, such as the Detroit Metro Area Communities Study (Wileden and Murphy, 2025) and the Baltimore Area Survey (Bader et al., 2025). Although these studies provide valuable insights, they capture transportation insecurity only within limited spatial and temporal contexts. This constraint restricts our ability to assess broader patterns, track changes over time, or make systematic comparisons across regions. As a result, transportation insecurity remains difficult to monitor at scale or on a sustained basis, leaving public agencies with limited empirical evidence to support program design, evaluation, and long-term planning.

In this project, we propose an alternative approach to measuring transportation insecurity (TI) by leveraging large-scale location intelligence data. Building on prior evidence that a traveler's TI status is closely associated with their mode use patterns (Wileden and Murphy, 2025), our approach infers TI through the extraction and interpretation of modal behavior observed in passively collected mobility data.

The core hypothesis is straightforward: if mode use patterns observed in TSI survey data are informative predictors of TI status, then analogous patterns obtained from mobile phone trajectory data can be used to infer the TI status of the corresponding device users. This insight forms the basis of a transfer learning framework. Specifically, we first train a TI classification model using labeled TSI survey data, and then apply the trained model to large-scale location intelligence datasets to infer TI patterns across a much broader population. By decoupling TI measurement from repeated survey administration, the transfer-learning approach offers a scalable pathway for monitoring transportation insecurity over time and across regions using existing mobility data sources.

A central pillar of our transfer-learning approach is the ability to infer reliable mode use patterns from location intelligence data. This requires, first, identifying the primary travel mode at the individual trip level and, second, aggregating trip-level classifications to the device level to characterize habitual mode use. Accomplishing this is nontrivial: location intelligence data are sparse, irregular in time, and lack ground-truth mode labels, making direct inference highly challenging.

To address this problem, we develop a semi-supervised machine learning framework that integrates trip attributes derived from location trajectories with mode-specific travel characteristics obtained from the Google Maps API. The framework proceeds in stages. We first identify a subset of "direct" trips whose observed travel time and distance closely match Google's suggested values for transit or driving, and use these matches to assign high-confidence initial labels. The validity of these labels is corroborated by examining the extent to which inferred transit and driving trajectories align with existing transit networks. These labeled trips are then used to train a transit–driving classifier that learns distinguishing features such as trip duration, speed, and path efficiency. When applied to the remaining unlabeled trips, the classifier identifies trips consistent with transit or driving while setting aside those that resemble neither. This residual group is analyzed using an exploratory hierarchical clustering procedure, which reveals distinct trip types, including active-mode trips and complex driving chains. Finally, we consolidate the inferred trip types into three primary modes—driving, transit, and active travel—and validate the resulting aggregate mode shares against those reported in regional

Travel surveys.

Our case study combines a location intelligence dataset capturing travel behavior in Chicago in May 2022 with a transportation insecurity (TI) classification model trained on the Detroit Metro Area Communities Study (DMACS) Wave 18 survey conducted in November 2023, which includes the Transportation Security Index (TSI). The results provide strong evidence for the feasibility and internal validity of the proposed methodology. The semi-supervised learning framework successfully distinguishes driving, transit, and active-mode trips from sparse mobile phone trajectories, and the resulting mode shares closely align with those reported in the Chicago Metropolitan Agency for Planning travel survey.

Using inputs from the DMACS survey, the TI classification model achieves approximately 70% accuracy and recall, even when mode use information is simplified to three broad categories and census-tract-level socio-demographic aggregates are used in place of individual characteristics. This indicates that predictive performance is only modestly reduced when adapting the model to features derived from location intelligence data. While directly transferring the Detroit-trained classifier to Chicago leads to overestimation of TI due to structural differences between the two regions, our preliminary analysis nonetheless identifies concentrations of transportation insecurity in census tracts consistent with well-documented spatial patterns of mobility disadvantage. These findings highlight both the current limitations and the substantial promise of the proposed transfer learning framework for measuring transportation insecurity at scale.

The remainder of this report is organized as follows. Section 2 reviews the background on transportation insecurity and the technical foundations of location intelligence data analytics. Section 3 presents the proposed transfer-learning framework and its key components, while Section 4 reports and discusses the empirical results. Finally, Section 5 concludes the report and highlights directions for future work.

2 Background

In this section, we first review the background of transportation insecurity in Section 2.1, and then turn to the key techniques in location intelligence data analytics that underpin our transfer learning framework in Section 2.2.

2.1 Transportation insecurity

2.1.1 Development of the Transportation Security Index (TSI)

Researchers and practitioners have traditionally relied on measures of material hardship to characterize the challenges faced by different segments of the population and to examine their associations with socio-demographic factors (Murphy et al., 2025). In transportation research, such challenges are often represented using demand- and supply-side indicators, including car ownership status (Smart et al., 2015) and composite accessibility measures (Grengs, 2012; Lind et al., 2025). While informative, these indicators do not provide a general or holistic account of how individuals actually *experience* transportation in their daily lives.

Motivated by insights from in-depth ethnographic studies and interviews with individuals facing substantial transportation-related difficulties, Gould-Werth et al. (2018) attempted to develop and validate a Transportation Security Index (TSI) designed to capture the lived symptoms of transportation insecurity (TI). Beginning with a set of 23 candidate indicators, they conducted an exploratory factor analysis on a pilot dataset and derived a 16-item questionnaire (TSI-16) that captures both material and relational manifestations of TI.

Subsequent work refined and strengthened the TSI-16 by improving internal consistency, clarifying the interpretation of TI categories, and increasing the parsimony of the instrument. Using a larger, nationally representative dataset, Murphy et al. (2021) demonstrated through confirmatory factor analysis that the material and relational dimensions of TI are empirically intertwined. They also standardized response scales across all survey items (never: 0, sometimes: 1, often: 2) to enhance consistency. The total score across items (ranging from 0 to 32) was used to represent the severity of transportation insecurity. Descriptive analyses indicated that approximately 18% of the U.S. population experiences TI, with higher prevalence associated with lower income, lack of car ownership, urban residence, and certain racial and ethnic groups (Murphy et al., 2022).

To further distinguish degrees of transportation insecurity, McDonald-Lopez et al. (2023) applied a clustering approach and identified five categories: transportation secure (0–2), marginal insecurity (3–5), low insecurity (6–10), moderate insecurity (11–16), and high insecurity (17 or above). More recently, the TSI-16 was abbreviated to a six-item instrument (TSI-6) by retaining items with the greatest response variability, and the five-category classification was consolidated into three groups: secure (0–1), marginal/low insecurity (2–5), and moderate/high insecurity (6–12) (Murphy et al., 2024). Using a split-ballot survey design, Murphy et al. (2024) showed that the TSI-6 performs comparably to the original TSI-16, estimating that approximately 17% of U.S. adults experience transportation insecurity.

The TSI-6 questionnaire consists of the following items:

1. In the past 30 days, how often did you have to reschedule an appointment because of a problem with transportation?

2. In the past 30 days, how often did you skip going somewhere because of a problem with transportation?
3. In the past 30 days, how often were you not able to leave the house when you wanted to because of a problem with transportation?
4. In the past 30 days, how often did you feel bad because you did not have the transportation you needed?
5. In the past 30 days, how often did you worry about inconveniencing friends, family, or neighbors because you needed help with transportation?
6. In the past 30 days, how often did problems with transportation affect your relationships with others?

2.1.2 Applications of TI

Transportation insecurity (TI) is not only associated with the socio-demographic characteristics of TSI survey respondents (Murphy et al., 2022), but is also linked to a broader set of transportation-related and non-transportation attributes. It is therefore natural to examine correlations between TI and other respondent characteristics captured within the same surveys. Existing evidence suggests that TI is related to a range of health and well-being outcomes, including increased depressive symptoms (McDonald-Lopez et al., 2023). TI has also been shown to co-occur with other forms of material hardship, such as food insecurity, housing insecurity, unmet medical needs, and difficulty paying household bills (Murphy et al., 2025).

Beyond nationally representative surveys, researchers have incorporated the TSI-6 and TSI-16 into regional surveys to support more localized, in-depth community analyses. In Baltimore, residents living in neighborhoods along the proposed Red Line corridor were found to be more likely to experience TI than residents elsewhere in the region (Bader et al., 2025). In Detroit, the share of the population experiencing TI is approximately double the national average, and TI is significantly associated with daily mode-use patterns—specifically, individuals reporting either no regular travel mode or multiple daily modes exhibit higher levels of TI (Wileden and Murphy, 2025). In Clark County, Nevada, Phillips et al. (2025) administered the TSI-16 to approximately one thousand older adults and found TI to be significantly related to income, age, and disability status.

It is also worth noting that researchers have examined concepts closely related to TI without directly employing the TSI-16 or TSI-6 instruments. For example, the 2022 National Household Travel Survey (NHTS) captures discrepancies between individuals' desired participation in out-of-home activities and their actual ability to do so, reflecting a simi-lar underlying concern (Robbennolt et al., 2025). Singer and Martens (2023) developed a 12-item questionnaire to measure travel difficulties—including trip challenges, reliance on others, and forgone travel—aimed at identifying less severe but more prevalent transportation problems. Chang et al. (2025) constructed a transport disadvantage index based on responses to transportation-related questions and found that higher disadvantage is associated with lower usage of smartphone-based mobility services.

Health-focused surveys have also highlighted transportation-related barriers. The National Health Interview Survey (NHIS) asks whether respondents delayed medical care due to transportation difficulties, and prior studies have found such difficulties to be associated with reduced likelihood of cancer screening (Pohl et al., 2025). Similarly, a retrospective cohort study surveying parents during infant-care visits found that those reporting difficulty accessing doctors' offices or pharmacies experienced higher rates of incomplete well-child visits, greater acute care utilization, and increased hospitalization (Park et al., 2025). Finally, Wander et al. (2025) reviewed a range of concepts and practices related to Universal Basic Mobility (UBM), which emphasizes access to sufficient mobility to meet daily needs, and identified the mitigation of transportation insecurity as a central component of this framework.

2.2 Location intelligence data analytics

Over the past two decades, location intelligence data have grown rapidly in volume, spatial coverage, and precision. As a result, researchers and practitioners have progressively shifted from in-vehicle GPS transponder data to phone-based call detail records (CDR), and more recently to commercially integrated location intelligence datasets (Dai et al., 2025). Because location intelligence data primarily consist of device trajectories—time-stamped location points—their use requires additional inference to recover higher-level behavioral information. This includes spatial attributes (e.g., activity locations and trip origins and destinations), temporal attributes (e.g., activity and trip start and end times, as well as durations), and semantic attributes (e.g., activity types and travel modes) (Fu et al., 2025).

In this project, we infer individual-level mode use patterns from location intelligence data through a three-step process: (1) identifying activities and trips, (2) inferring the most likely travel mode for each trip, and (3) aggregating trip-level mode information to the device level.

Identifying activities and trips is relatively well established. Most studies rely on rule-based methods in which consecutive observations with minimal spatial displacement are classified as activities, while movements between activities are treated as trips (Ahas et al., 2010; Calabrese et al., 2013; Chen et al., 2014). A device's "home" location is typically inferred as the activity location occupying the majority of nighttime hours on weekdays

(Pappalardo et al., 2021). In our prior work, we followed this standard approach to identify activities and trips in the location intelligence data used in this project (Dai et al., 2025).

Inferring travel mode for each trip, however, is substantially more challenging. Limitations in spatial accuracy and temporal sampling frequency often make it difficult to learn modes directly from trajectories alone. Consequently, many studies augment location intelligence data with external information, such as transportation networks and web-based mapping services, to support mode inference (Huang et al., 2019). Because location intelligence data typically lack ground-truth mode labels, most approaches rely on unsupervised or probabilistic machine learning techniques.

For example, Chen et al. (2019) constructed features such as travel distance, speed, and ratios between observed speeds and map-based estimates for driving and walking, and then applied a fuzzy k-means algorithm to distinguish among bus, car, subway, walking, and biking. Bachir et al. (2019) focused on point-level features, first clustering trajectory segments with similar characteristics and assigning provisional mode labels, then propagating mode probabilities along trajectories using a Bayesian inference framework. Similarly, Liu et al. (2025) employed Bayesian inference and hidden Markov models (HMMs) to infer trip chains and capture multimodal travel behavior. These studies relied on transportation network data to construct mode-specific features and validated their results at the aggregate level using travel survey mode shares and public transit smart card data (Bachir et al., 2019; Liu et al., 2025). Zhong et al. (2024) identified bus trips by directly matching location intelligence data with fixed-route bus GPS trajectories. Across these studies, proximity to transportation infrastructure plays a central role in distinguishing travel modes.

3 Methodology

Figure 1 presents a high-level overview of the proposed transportation insecurity (TI) transfer learning framework. The workflow consists of three main stages. First, we process raw location intelligence data to identify individual activities and trips (Section 3.1), and subsequently infer travel modes for each trip in order to construct device-level mode use patterns (Section 3.2). These patterns serve as behavioral signatures that summarize how individuals move through the transportation system.

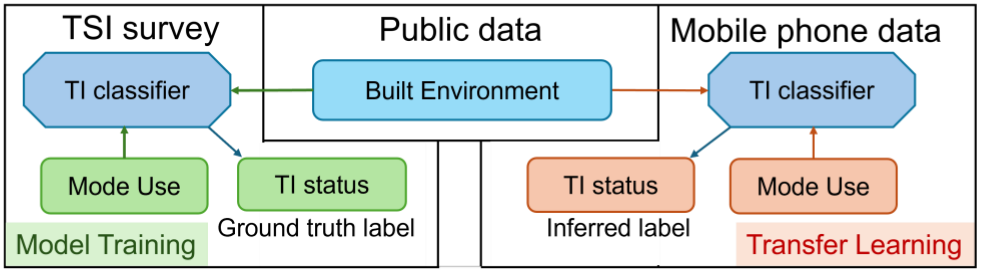

Figure 1: A transfer learning framework.

Second, we use a transportation security index (TSI) survey dataset to train a TI classification model that maps mode use patterns—together with other relevant covariates—to observed TI outcomes (Section 3.3). This step establishes the behavioral link between observed travel patterns and transportation insecurity as measured in survey data.

Finally, we transfer the trained classifier to the location intelligence dataset and apply it to the full population of observed devices to infer their TI status at scale (Section 3.4).

3.1 Activity and trip identification

Mobile phone location data consist of semi-continuous trajectories—time-stamped sequences of geographic coordinates associated with individual mobile devices. By systematically analyzing these trajectories, it is possible to decompose daily mobility into two fundamental components: activities, defined by sustained presence at a location, and trips, defined as movements connecting successive activities. This activity–trip representation provides a behavioral abstraction that is widely used in transportation research and forms the foundation of our analysis.

Following established practice in the literature (Ahas et al., 2010; Calabrese et al., 2013; Alexander et al., 2015; Chen et al., 2014; Fu et al., 2025), we infer activities based on both spatial and temporal persistence. Specifically, an activity is identified when a device remains within a confined spatial area for a non-trivial duration. In this study, we define an activity as a sequence of consecutive trajectory points whose locations remain effectively unchanged for at least 15 minutes. This threshold reflects a balance between filtering out transient stops (e.g., traffic signals or short interruptions) and retaining meaningful engagements with a location.

To account for spatial noise and minor positional variation in mobile phone data, we further cluster activities that occur in close proximity. Two activities are considered spatially equivalent if the distance between their geometric centers is within 200 meters, in which case they are assigned to the same unique clustered location (Hariharan and Toyama, 2004). This clustering step allows us to consolidate repeated visits to the same functional place—such as a residence, workplace, or transit hub—into a single location entity.

The home location of a device is identified as the unique clustered location at which the device spends the majority of its weekday nighttime hours, consistent with prior studies (Pappalardo et al., 2021). Activities occurring at this location are labeled as home activities. All remaining clustered locations are then ranked by total accumulated activity duration. The locations with the longest and second-longest durations are designated as the device's primary and secondary activity locations, respectively, which often correspond to work, school, or other regularly visited destinations.

Once activities are identified and labeled, trips are defined as the movements between consecutive activities in a device's trajectory. This activity-based decomposition enables subsequent inference of travel modes and the construction of individual-level mode use patterns, which are central inputs to our transportation insecurity transfer learning framework.

3.2 Travel mode inference

We adopt a semi-supervised labeling approach that combines rule-based weak supervision, supervised classification, and unsupervised clustering to expand and refine mode labels. To formally define our approach, let us first define a set of J trips identified from the location intelligence data and let $j \in J \equiv \{1, 2, \ldots, J\}$denote the indices of trips. A selected set of key attributes of each the trip j are defined as follows:

1. Duration t_j: the time difference between the beginning and end of trip j;
2. Distance d_j: the total travel distance of trip j, calculated by summing the distances between consecutive data points included in trip j;
3. Straight line distance l_j: the haversine distance between the beginning and end locations of trip j;
4. Average speed v_j: the ratio of d_j and t_j;
5. Weekday indicator w_j: 1 if trip j begins on a weekday, 0 otherwise;
6. Peak hour indicator k_j: 1 if trip j begins within the peak hour window—6-10am and 4-8pm, 0 otherwise.

The goal of mode inference is to assign to each trip j a mode $m_j \in M \equiv \{p, c, a\}$, where p denotes public transit, c is driving and a is active modes.

3.2.1 Initial labels

The first step is to identify a subset of trips whose observed travel time and distance closely align with the mode-specific travel time and distance suggested by Google Maps for transit or driving. The underlying premise is that, although Google Maps recommendations do not capture the full complexity of real-world travel behavior—such as detours, intermediate stops, or mixed-mode trips—there exists a subset of *direct* trips in which travelers closely follow the fastest routes recommended by navigation services. These trips can be reliably distinguished as either transit or driving and therefore serve as high-confidence reference examples.

By exploiting this property, we assign initial mode labels to trips that exhibit strong consistency with Google Maps benchmarks, effectively creating a weakly supervised training set. These initial labels provide an empirical anchor for subsequent learning steps, allowing mode-specific characteristics to be inferred directly from the data rather than imposed through strong behavioral assumptions.

Mode-specific reference travel times and distances between origins and destinations are obtained through Google Maps API queries. To reduce computational cost and ensure scalability, we use the centroids of the geographic units (e.g., census tracts or community areas) containing the actual trip origins and destinations as spatial proxies. This approximation preserves the large-scale spatial structure of trips while substantially limiting the number of API queries required, making the approach feasible for large datasets.

We denote the Google map reference time and distance of trip j for mode $m \in \{p, c\}$as t_j^m and d_j^m, respectively. To identify the direct transit and driving trips, we apply a range criterion, which assigns p or d to initial mode label $m_j^L \in \{p, c, n\}$for all trips, where n denotes "not direct transit or driving trip":

$$m_j^L = \begin{cases} p, \text{if } \left(\frac{\bar{t}_j^{\,p}}{t_j}, \frac{\bar{d}_j^{\,p}}{d_j} \right) \in \left[1 - \alpha_p, 1 + \alpha_p\right]^2; \\ d, \text{if } \left(\frac{\bar{t}_j^{\,c}}{t_j}, \frac{\bar{d}_j^{\,c}}{d_j} \right) \in [1 - \alpha_c, 1 + \alpha_c]^2; \\ \qquad n, \text{otherwise} \end{cases}, \forall j \in J,$$

where α_p and α_c are range tolerance parameters for transit and driving respectively. We further define the initially labeled trip set $J^L \equiv \{j \in J \mid m_j^L \in \{p, c\}\}$, which serve as a subset of trips with ground truth mode labels. For the ease of reference, we also define the initially labeled transit and driving trip sets as $J_p^L \equiv \{j \in J \mid m_j^L = p\}$ and $J_c^L \equiv \{j \in J \mid m_j^L = c\}$ respectively.

3.2.2 Semi-supervised classification

Semi-supervised learning (also referred to as weak supervision) aims to infer labels for a large volume of unlabeled data by leveraging a relatively small labeled subset from the same dataset, under the assumption that labeled and unlabeled instances are drawn from similar feature distributions (Zhuang et al., 2020). This paradigm is particularly well suited to location intelligence data, where obtaining reliable ground-truth labels at scale is costly or infeasible.

In this step, we exploit the initially labeled trip set J^L—consisting of high-confidence transit and driving trips identified through map-based weak supervision—to learn the distinguishing characteristics of these two modes. Using the trip-level feature vectors, we train a supervised classifier that captures systematic differences in travel time, distance, speed, and temporal patterns between transit and driving trips. The trained classifier is then applied to the remaining unlabeled trip set $J^U \equiv J \backslash J^L$, allowing us to propagate mode labels beyond the small reference set while maintaining consistency with observed data patterns.

First, we train a supervised transit–driving classifier, denoted by $f_{p/c}(\cdot)$, using the initially labeled trip set J^L. The classifier takes as input the trip-level feature vector $\boldsymbol{x}_j = (t_j, d_j, l_j, v_j, w_j, k_j)$ and outputs a predicted mode label p (public transit) or c (driving). Model training follows a standard k-fold cross-validation procedure to ensure robustness and to mitigate overfitting, given the limited size of the labeled subset.

Next, we apply the trained classifier $f_{p/c}(\cdot)$ to all unlabeled trips $j \in J^U$. Because $f_{p/c}(\cdot)$ is designed to discriminate only between transit and driving, trips belonging to other modes (e.g., walking or cycling) would be forcibly assigned to one of these two classes if no additional screening were applied. To prevent such misclassification, we introduce a distance-based filtering step that identifies trips that are dissimilar to both transit and driving patterns.

Specifically, we compute the Mahalanobis distance[1] between the feature vector $\boldsymbol{x}_j$ of each unlabeled trip and the class centers of the initially labeled transit and driving trips. This allows us to assess whether a trip plausibly belongs to either class based on its multivariate similarity, before assigning a definitive mode label. The means and covariance matrices of the initially labeled transit and driving trips are defined as follows:

Mode use patterns are measured differently in survey data and trajectory data,

[1] The Mahalanobis distance measures the distance between a point and a distribution while accounting for scale and correlation among variables.

$$\mu_p = \frac{1}{\left|\mathbb{J}_p^L\right|}\sum_{j\in\mathbb{J}_p^L} \boldsymbol{x}_j\,, \quad \Sigma_p = \frac{1}{\left|\mathbb{J}_p^L\right|-1}\sum_{j\in\mathbb{J}_p^L}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_p\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_p\right)^{\top};$$

$$\mu_c = \frac{1}{\left|\mathrm{J}_c^L\right|}\sum_{j\in\mathrm{J}_c^L} \boldsymbol{x}_j\,, \quad \Sigma_c = \frac{1}{\left|\mathrm{J}_c^L\right|-1}\sum_{j\in\mathrm{J}_c^L}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_c\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_c\right)^{\top}.$$

And the Mahalanobis distances between $\boldsymbol{x_j}$ and class centers ($\boldsymbol{\mu}_p$ and $\boldsymbol{\mu}_c$) are:

$$r_p\left(x_j;\, \mu_p\right) \;=\; \sqrt{\left(\left(\boldsymbol{x}_j \,-\, \boldsymbol{\mu}_p\right)^{T} \Sigma_p^{-1}\left(\boldsymbol{x}_j \,-\, \boldsymbol{\mu}_p\right)\right)};$$

$$r_c\left(x_j;\, \mu_c\right) \;=\; \sqrt{\left(\left(\boldsymbol{x}_j \,- \boldsymbol{\mu}_c\,\right)^{T} \Sigma_c^{-1}\left(\boldsymbol{x}_j \,-\, \boldsymbol{\mu}_c\right)\right)}.$$

Now we can properly assign a mode label $m_U \in \{p, c, o\}$, where o stands for “others”, to each of the unlabeled trips by applying Mahalanobis distance filter:

$$m_j^U \;=\; \begin{cases} o, \text{if } r_p\left(\boldsymbol{x}_j, \boldsymbol{\mu}_p\right) > \hat{r}_p \text{ and } r_p\left(\boldsymbol{x}_j, \boldsymbol{\mu}_c\right) > \hat{r}_c; \\ \qquad f_{p/c}(\boldsymbol{x}_j), \text{otherwise} \end{cases} ,\forall\, j \;\in\; J^U,$$

where $\hat{r}_p$ and $\hat{r}_c$ are the distance thresholds to the transit and driving class center. The mode o includes a variety of trips that do not resemble normal transit or driving trips, and we will further investigate the mixture in the next part.

3.2.3 Unsupervised clustering

The final step of the mode inference procedure applies an unsupervised clustering algorithm to the set of trips labeled as o, which includes trips that do not resemble typical transit or driving patterns. The goal of this step is to uncover latent structure within this heterogeneous set and to identify distinct travel modes based on their observed attributes. We expect that these trips will naturally separate into multiple clusters, each corresponding to a different mode or subtype of travel behavior.

To preserve transparency and interpretability, we employ an agglomerative hierarchical clustering algorithm (HCA), which we have previously applied to the same dataset to identify different types of workers (Dai et al., 2025). In this study, HCA begins by treating each trip as its own cluster and iteratively merges the pair of clusters with the smallest distance according to a predefined linkage criterion. This process continues until all trips are merged into a single cluster, producing a dendrogram that reveals the hierarchical structure of the data and the sequence in which clusters are formed. An important advantage of HCA is that it does not require the number of clusters to be specified a priori; instead, we determine the appropriate level of aggregation by examining cluster characteristics and separation.

We expect that active modes of travel (e.g., walking or cycling) will emerge as clusters characterized by shorter distances and lower speeds. Additional clusters may capture atypical or mixed forms of driving and transit trips. Because the precise cluster structure is data-dependent, a detailed discussion of the identified clusters is deferred to Section 4. After clustering, we consolidate the results by merging all transit-related clusters and all driving-related clusters, and retain active modes as a separate category. This yields the final mode label $m_j \in \{p, c, a\}$for each trip $j \in J$.

To validate the inferred mode labels, we examine whether trips classified as transit exhibit spatial alignment with the transit network, and we compare the aggregate mode shares derived from our data with those reported in regional travel surveys. Together, these checks provide external validation of the plausibility and consistency of the inferred travel modes.

3.3 Learning transportation insecurity

In this section, we develop a supervised machine learning framework to predict transportation insecurity (TI) using survey respondents' socio-demographic characteristics and travel behavior. Let I denote the total number of survey respondents, indexed by $i \in I \equiv 1, 2, \dots, I$. Our modeling approach draws on variables commonly available in TSI surveys, complemented by publicly accessible contextual data. We organize the explanatory variables into the following categories:

1. Individual level socio-demographic variables $\boldsymbol{y}_i^{SD}$: A binary vector indicator capturing gender, race, age, income, etc. For example, if the variable set is {male, female, white, non-white}, a black male respondent will be recorded as (1, 0, 0, 1).
2. Census tract level aggregate socio-demographic variables $\bar{\boldsymbol{y}}_i^{SD}$: A vector of the percentages of different socio-demographic groups in the home census tract for respondent i.
3. Census tract level living environment variables $\bar{\boldsymbol{y}}_i^{LE}$: A vector of variables that describe the local living environment of the home census tract of respondent i, such as transit frequency, accessibility, etc.
4. Individual level mode use pattern variables $\boldsymbol{y}_i^{m}$: A vector of variables that describe different mode use frequencies (probability of using a mode on a typical day) and overall mode availability and usage of respondent i.

These variables capture both individual circumstances and contextual factors that are hypothesized to shape transportation insecurity, and they form the input feature space for the supervised learning model developed in the following.

The transportation insecurity (TI) status of each survey respondent is determined based on their responses to the TSI-6 questionnaire, following the categorization scheme proposed by Murphy et al. (2024). For analytical tractability and to align with the needs of downstream transfer learning, we further simplify these categories into a binary outcome: a respondent is classified as either transportation secure (0) or transportation insecure (1). Accordingly, the TI label for respondent i is denoted by $TI_i \in \{0, 1\}$.

Our objective is to learn a TI classifier $f_{TI}(\cdot)$ that maps combinations of individual and contextual attributes—specifically $\boldsymbol{y}_i^{SD}$, $\overline{\boldsymbol{y}}_i^{LE}$ and $\boldsymbol{y}_i^{m}$ —to a predicted TI status $\widehat{TI}_i \in \{0, 1\}$. A key design requirement is that $f_{TI}(\cdot)$ be compatible with inputs derived from location intelligence data. Unlike survey data, location intelligence data do not contain individual-level socio-demographic information $\boldsymbol{y}_i^{SD}$. Therefore, the classifier must also perform well when individual-level attributes are replaced by census-tract-level aggregates $\overline{\boldsymbol{y}}_i^{SD}$, yielding a tract-based TI prediction $\widehat{TI}_i \in \{0, 1\}$.

Given the policy relevance of correctly identifying transportation-insecure individuals, we prioritize performance on both accuracy and recall. In particular, recall is critical to minimizing false negatives, which would otherwise understate the prevalence of transportation insecurity. Following standard supervised learning practice, we evaluate out-of-sample performance using a held-out test set indexed by I_{test}, and report both accuracy and recall based on the predicted outcomes.

$$\widehat{Acc} = \frac{\sum_{i \in I_{test}} \mathbf{1}\left[\widehat{TI}_i = TI_i\right]}{|I_{test}|}, \widehat{Rec} = \frac{\sum_{i \in I_{test}} \mathbf{1}\left[\widehat{TI}_i = TI_i = 1\right]}{\sum_{i \in I_{test}} \mathbf{1}[TI_i = 1]};$$

$$\overline{Acc} = \frac{\sum_{i \in I_{test}} \mathbf{1}[\overline{TI}_i = TI_i]}{|I_{test}|}, \overline{Rec} = \frac{\sum_{i \in I_{test}} \mathbf{1}[\overline{TI}_i = TI_i = 1]}{\sum_{i \in I_{test}} \mathbf{1}[TI_i = 1]}.$$

where $\mathbf{1}[\cdot]$ is the binary indicator function, $\widehat{Acc}$ and $\widehat{Rec}$are the accuracy and recall based on $\widehat{TI}_i$, and $\overline{Acc}$ and $\overline{Rec}$ are the accuracy and recall based on $\overline{TI}_i$.

3.4 Transfer learning

Transfer learning refers to the practice of leveraging knowledge acquired from a learning task in a *source domain* to improve performance on a related task in a *target domain*, particularly when the two domains and/or tasks are not identical but share underlying structure[2] (Pan and Yang, 2009; Zhuang et al., 2020). The central premise is that certain representations, relationships, or decision boundaries learned from one data-generating process can be reused—often with adaptation—to inform inference in another.

In our context, the source domain consists of TSI survey data, where transportation insecurity labels are directly observed and mode use patterns are self-reported. The target

[2] If both the domains and the tasks are identical; this setting reduces to semi-supervised learning.

domain is location intelligence data, where mode use patterns must be inferred from trajectories and no ground-truth TI labels are available. Applying the trained classifier $f_{TI}(\cdot)$ across these domains presents two distinct challenges. First, there is a *domain shift*: mode use patterns are measured differently in survey data and trajectory data, leading to discrepancies in feature definitions and distributions. Second, there is a potential *task shift*: even when observable features coincide, the relationship between those features and TI status may vary across social, spatial, or institutional contexts.

In this study, we focus on addressing the domain shift by carefully reconciling and harmonizing feature representations across the two data sources. Specifically, we construct mode use variables from location intelligence data that are conceptually aligned with those used in the TSI surveys, enabling the learned decision function to generalize across domains. Addressing task shift—such as context-dependent differences in how similar travel behaviors map to transportation insecurity—would require additional labeled data or contextual modeling and is therefore left for future research.

4 Results

To evaluate the proposed transfer-learning framework, we conduct a case study that combines two independent data sources: a large-scale location intelligence dataset capturing activity and travel patterns of residents in the Chicago metropolitan area in May 2022, and a Transportation Security Index (TSI) survey administered to residents of the Detroit metropolitan area in November 2023. Although these datasets differ in geographic context, time period, and data-generating process, this setting provides a meaningful test of the framework's feasibility. Specifically, it allows us to examine whether a classifier trained on survey-based measures of transportation insecurity can be successfully transferred to infer population-level transportation insecurity from passively collected mobility data. The results presented below therefore focus on demonstrating the robustness and practical viability of the proposed approach, rather than on drawing definitive conclusions about transportation insecurity levels in any specific region.

4.1 Data

The location intelligence data used in this study consist of anonymized mobile phone records acquired from PickWell Sp. z.o.o., a European data provider, and subsequently preprocessed by Oplytix, LLC, a U.S.-based consulting firm. The raw dataset contains only time-stamped geographic coordinates for individual mobile devices, without any explicit information on activities or trips. We therefore first infer activities and trips using an algorithmic framework that was developed in this study and has been published in Dai et al. (2025).

Following established practice, we treat an activity as occurring when an individual remains at approximately the same location for a nontrivial duration. Specifically, an activity is defined as a consecutive sequence of trajectory points with no significant spatial displacement that persists beyond a dwell-time threshold of 15 minutes. To characterize regular activity patterns, we identify a device's "home" location as the unique clustered location that accounts for the majority of its weekday nighttime activity, consistent with prior vali-dation studies (Pappalardo et al., 2021).

From the raw mobile phone records, we construct a device-level dataset in which each device is represented by a unique anonymized identifier (a pseudo ID). Each record is classified into one of four event types: activity, trip, short-sleep, and long-sleep. Our analysis focuses on activities and trips, which together capture the spatial–temporal structure of daily mobility.

For each activity, we extract its duration and assign its location to a census tract. To protect user privacy, precise longitude and latitude information is not available; instead, all events are spatially aggregated to census tracts and encoded using the corresponding GEOID (City of Chicago, 2024).

For each trip, we compute key attributes including travel distance, duration, and average speed, and assign both the origin and destination to census tracts. In addition, we construct a CTA rail alignment feature by checking, at each timestamp, whether the device location falls within a 50 m buffer of the CTA rail network. By aggregating this indicator over the full trip sequence, we estimate the distance traveled and time spent along rail-aligned segments, which later serves as an important signal for transit mode inference.

Based on activity patterns, we identify a set of frequently visited, non-home clustered locations for each device and rank these locations by total dwell time, following the procedure described in Section 3.1. By convention, the home location is assigned rank 1, while the non-home location with the highest cumulative activity duration (rank 2) is commonly interpreted as a proxy for the work location. Figure 2 maps the spatial distribution of inferred rank-group 1 (left) and rank-group 2 (right) activity locations across Chicago census tracts. The right map shows a pronounced rank-group 2 activity concentration in and around the Loop, consistent with the city's central business district (CBD) and its high density of jobs. The left map shows that devices' home locations are more broadly distributed across peripheral tracts, a direct contrast with the right map. We note that these spatial patterns reflect the observed device sample and the clustering-based inference procedure; absolute intensities may therefore be influenced by sampling coverage, filtering choices, and location assignment uncertainty.

In May 2022, the mobile phone dataset contains 121,754 unique devices, comprising

7,178,028 trip records and 6,058,291 activity records. After applying the home–census-tract identification algorithm described in Section 3.1, we infer that 52,799 devices are associated with Chicago residents, defined as devices whose inferred home census tract lies within the City of Chicago. These Chicago-resident devices account for 3,764,502 trip records and 3,379,875 activity records.

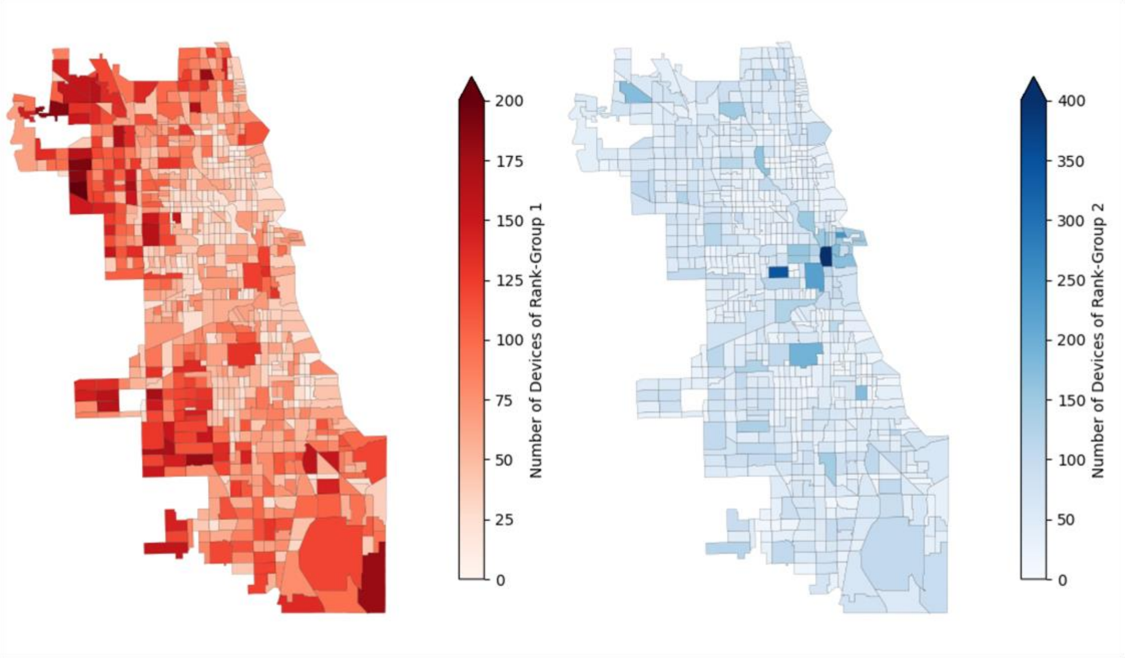
200
175
150
125
100
75
50
25
0
Number of Devices of Rank-Group 1
400
350
300
250
200
150
100
50
0
Number of Devices of Rank-Group 2

Figure 2: Heatmap of number of devices in rank-group 1 and rank-group 2 in May 2022.

To better characterize routine mobility patterns within the city and to limit the influence of long-distance or incidental travel by nonresidents, we restrict all subsequent analyses to this Chicago-resident subset. On average, each inferred Chicago-resident device generates 71.30 trips and 64.01 activities over the study period. Figure 3 summarizes the distributions of per-device trip and activity counts.

Consistent with expectations, trip counts tend to exceed activity counts at the device level, reflecting the fact that individuals typically undertake multiple trips between successive activity episodes. This pattern provides a useful validation of the activity–trip identification procedure and underscores the richness of the inferred mobility sequences used in subsequent analyses.

Figure 4 presents the distributions of inferred trips and activities by day of week and time of day. Overall, trip counts exceed activity counts across most temporal bins. Activity and trip frequencies are relatively stable across weekdays, with a modest decline observed on Thursdays.

Clear diurnal patterns emerge in the time-of-day distributions. Trips occur predominantly during daytime hours, reflecting routine commuting and daily travel, while both trips and activities are sparse between 12 AM and 5 AM. During these overnight hours, activity counts slightly exceed trip counts, consistent with extended stationary periods such as night-time rest. The close alignment between trip and activity temporal profiles provides further reassurance that the activity–trip segmentation captures meaningful behavioral structure.

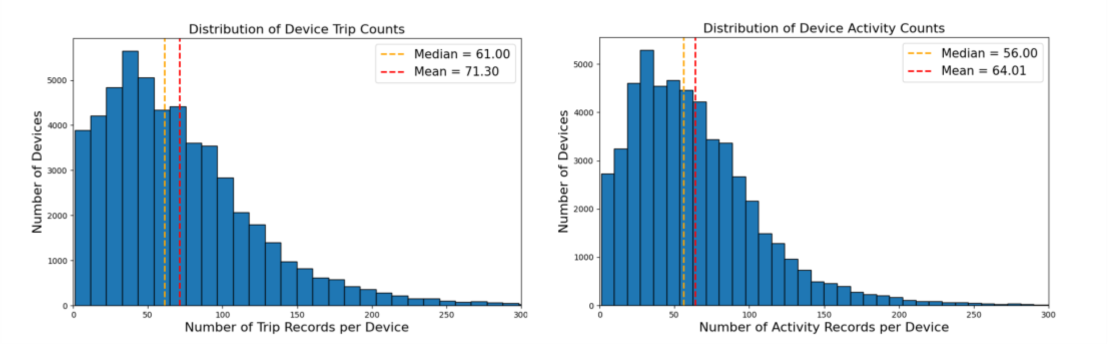

Figure 3: Trip and activity count distribution per device in May 2022.

Table 1 reports descriptive statistics for inferred activities and trips. Both exhibit substantial heterogeneity across records. Activity durations are concentrated at short to moderate lengths but display a pronounced right tail corresponding to prolonged stays, such as work or home activities. Trips, by contrast, are typically short in both duration and distance, yet include occasional long-distance movements, highlighting the diversity of travel behavior captured in the dataset.

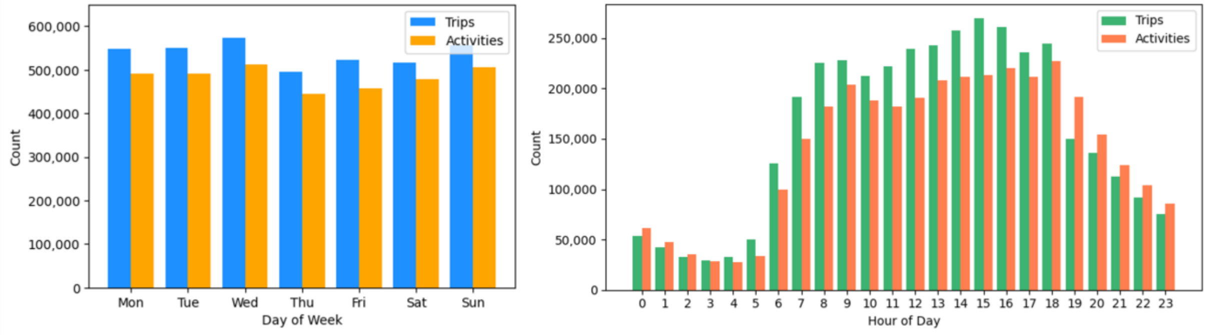

Figure 4: Activity and trip count temporal distribution in May 2022.

4.2 Mode use patterns in Chicago

Due to sampling inconsistency, many devices do not provide observations spanning the full study period or contain trips with insufficient information for reliable inference. To ensure data quality, we impose a series of filters at both the device and trip levels.

First, we require that a device have trajectory data covering at least 26 of the 28 days

Table 1: Descriptive statistics for recorded activities and trips.

	Activity duration	Trip duration	Trip distance	Average speed	Straight-line distance
	hr	hr	km	km/hr	km
mean	6.78	0.62	13.26	27.29	4.70
25%	0.64	0.14	1.46	7.50	0
50%	1.86	0.35	5.67	16.26	1.14
75%	8.47	0.75	15.53	28.88	5.82
std	15.38	0.94	24.76	157.31	8.09

in the study period. This criterion helps ensure that the observed travel patterns are representative of regular behaviors rather than sporadic or incidental activities.

Second, we apply a set of filters to the trips associated with the remaining devices to exclude observations unsuitable for the mode inference algorithm. Specifically, each retained trip must satisfy the following conditions: (i) contain more than six trajectory points; (ii) have a duration $t_j \in [0.05, 2]$ hours; (iii) have a distance $d_j \in [0.5, 40]$ km; and (iv) have an average speed $v_j \in [1, 100]$ km/hour. These thresholds remove implausible movements, extremely short displacements, and poorly sampled trajectories.

After applying these filters, the dataset is reduced to 2,624 *study devices*, comprising a total of 87,074 *proper trips*, which form the basis for subsequent mode inference and analysis.

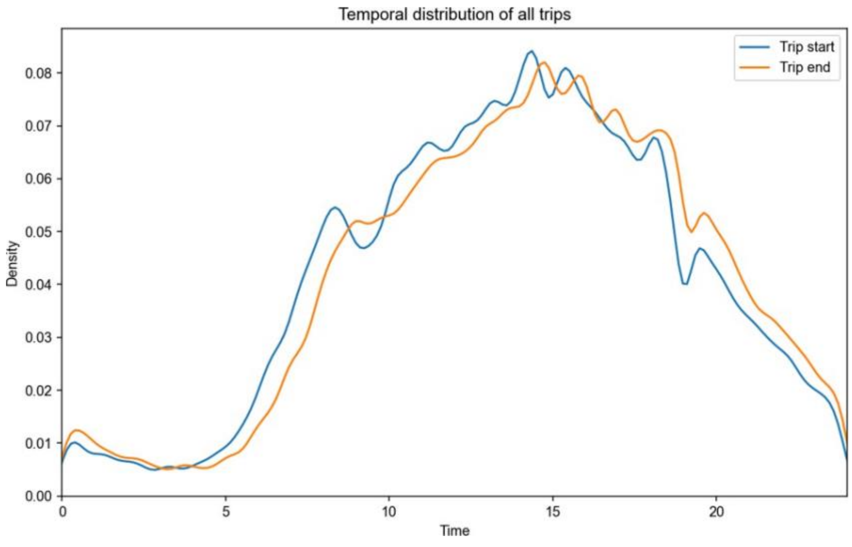

Figure 5: Temporal distribution of all proper trips.

Figure 5 illustrates the temporal distribution of trip start and end times for all *proper trips* included in the analysis. As expected, the majority of trips occur during daytime hours, with pronounced peaks corresponding to typical morning and evening commute periods.

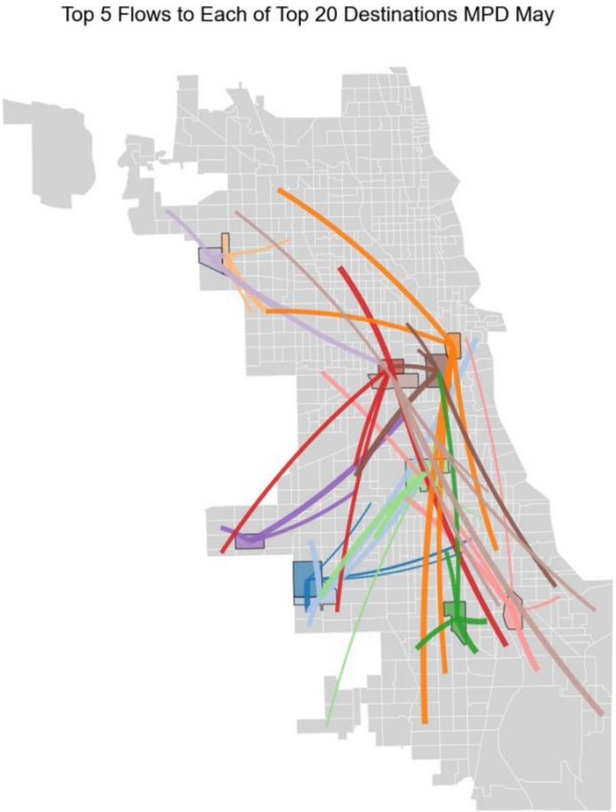

Figure 6: Top origin-destination pairs in all proper trips.

Notably, a substantial volume of trips is also observed during the afternoon hours, suggesting heterogeneous travel purposes beyond standard work-related commuting.

Figure 6 displays the top 20 destination census tracts, with color intensity indicating destination frequency, along with the top five origin–destination flows into each major destination, shown as curved links in the corresponding destination color. Several prominent activity centers are clearly identifiable, including the Chicago Loop (central business district), highlighted in orange, and the University of Illinois Chicago (UIC) area, highlighted in brown. In addition, a noticeable dispersion of popular destinations appears across sub-urban tracts, further underscoring the diversity of regional travel patterns captured in the data.

We next apply the mode inference procedure described in Section 3.2 to the 87,074 proper trips. Using tolerance parameters $\alpha_p = 0.15$ and $\alpha_c = 0.2$, we identify 6,888 trips whose observed travel time and distance closely match Google Maps reference values for either transit or driving. This labeled subset J^L consists of 5,098 driving trips and 1,790 transit trips.

To assess the plausibility of these initial labels, we examine the extent to which labeled trips align spatially with the Chicago Transit Authority (CTA) rail network, which is largely segregated from the roadway system[3]. We find that 52% of the initially labeled transit trips contain at least one trajectory point aligned with CTA rail routes, compared to only 41% of labeled driving trips. Moreover, transit trips spend approximately 33% of their total travel distance and 26% of their total travel time along rail-aligned segments, substantially higher than the corresponding figures for driving trips (25% of distance and 19% of time). These differences provide supporting evidence that the initial labeling procedure captures meaningful modal distinctions.

Using the labeled set J^L, we then train a supervised transit/driving classifier $f_{p/c}(\cdot)$ based on the full feature set: travel duration, travel distance, average speed, straight-line distance, weekday indicator, and peak-hour indicator. Implemented as a random forest model, the classifier achieves an out-of-sample accuracy of 96%, indicating that the selected trip attributes strongly differentiate transit and driving behaviors. Figure 7 reports feature importance rankings, showing that average speed and travel duration are the two most influential predictors, while the remaining features contribute relatively less to distinguishing between these two modes for the initially labeled trips.

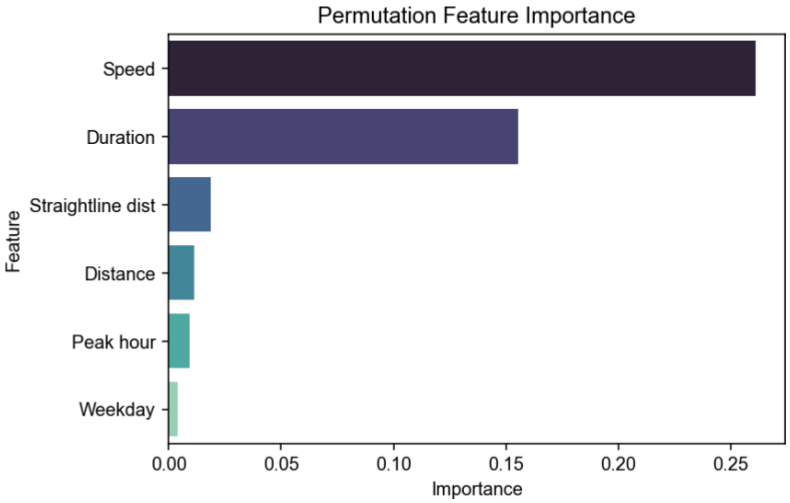

Figure 7: Feature Importance ranking for the transit/driving classifier.

Next, we assign mode labels to all previously unlabeled trips in the set J^U using the rule defined in Equation (4). A key step in this process is determining the Mahalanobis distance thresholds that identify trips which resemble neither typical transit nor driving behavior. To this end, we compute the empirical distribution of Mahalanobis distances from each unlabeled trip to the centroids of the labeled transit and driving classes. We then set the distance thresholds at the 85th percentile of these distributions. Intuitively, trips

[3] Notable exceptions include portions of the Blue Line (overlapping with I-90 and I-290), the Red Line (overlapping with I-90/94 in the south), and segments within the Loop.

that fall within the top 15% of distances from *both* class centroids are considered sufficiently dissimilar from standard transit and driving trips and are therefore labeled as *others*.

This filtering step serves two purposes. First, it prevents the forced assignment of clearly atypical trips—such as short active-mode trips or complex trip chains—to inappropriate motorized modes. Second, it preserves the integrity of the transit and driving classes by ensuring that only trips with reasonable similarity to the labeled examples are classified as such. By combining the initial labels m_j^L for trips in J^L with the inferred labels m_j^U for trips in J^U, we obtain a complete preliminary mode labeling for all trips in J.

Table 2: Conditional feature means for different preliminary mode classes.

Mode	No. of trips	$\bar{t}$ (h)	d (km)	l (km)	$\bar{v}$ (km/h)	k	$\bar{w}$
Driving	29,269	**0.38**	10.7	7.9	**22.7**	0.43	0.76
Transit	22,754	**0.61**	8.4	5.2	**12.7**	0.44	0.79
Other	35,051	0.76	16.3	5.7	23.7	0.38	0.70

Table 2 reports the conditional means of the six trip-level predictors for each preliminary mode class, with the key discriminating predictors highlighted in bold. Clear and intuitive differences emerge across the classes. Driving trips are characterized by substantially higher average speeds and shorter travel durations than transit trips, reflecting the greater flexibility and door-to-door nature of automobile travel. Transit trips, while slower on average, tend to span longer durations, consistent with access, waiting, and transfer components embedded in public transportation use.

Both driving and transit trips exhibit markedly higher proportions of weekday and peak-hour travel relative to trips classified as *others*, underscoring the central role these motorized modes play in routine commuting and work-related mobility in Chicago. In contrast, trips labeled as *others*—which include active modes and atypical travel patterns—are less concentrated during peak periods and weekdays, suggesting greater association with discretionary, local, or non-work activities. Taken together, these patterns provide further validation that the preliminary labeling captures meaningful behavioral distinctions across travel modes.

Next, we conduct an exploratory clustering analysis on the trips labeled as “other” to further investigate travel behaviors that differ substantially from direct driving and transit trips. Using the same set of trip-level features, we apply hierarchical agglomerative clustering (HCA) to uncover latent structure within this heterogeneous group. Figure 8 presents the resulting dendrogram, which visualizes the cluster formation process. The horizontal axis lists all “other” trips, ordered by similarity, while the vertical axis reports the within-cluster

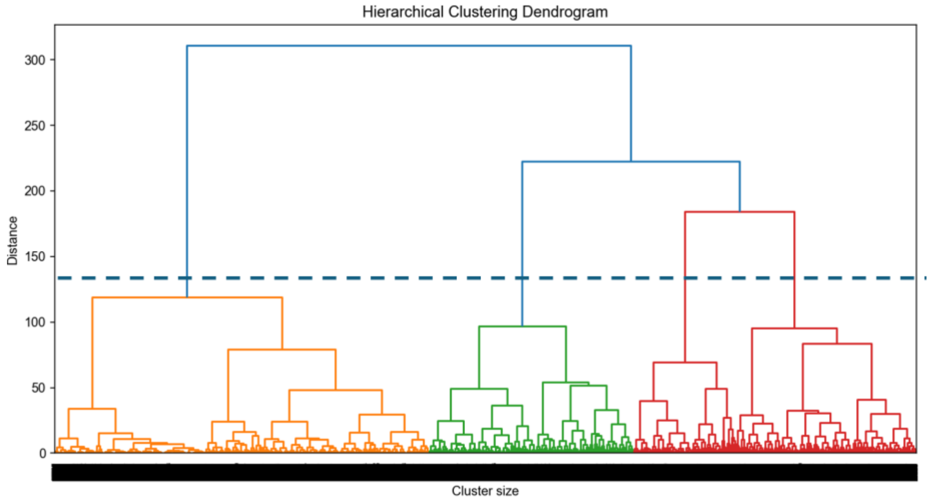

Figure 8: HCA results on other trips.

dissimilarity at each stage of the hierarchy. The height of each vertical merge therefore represents the loss in within-cluster similarity incurred when two clusters are combined.

A common heuristic for selecting the number of clusters is to identify a horizontal cut that intersects long vertical segments, indicating that additional merging would substantially degrade cluster coherence. Applying this criterion, we select a cutoff at a distance of approximately 130, yielding four dominant clusters. The conditional means of the key trip attributes for these clusters are reported in Table 3.

Based on their distinct behavioral signatures, we label the four clusters as *active modes*, *long driving*, *urban trip chains*, and *suburban trip chains*. Trips in the active-modes cluster exhibit slow speeds, short travel distances, and short straight-line distances, consistent with walking and bicycling behavior[4]. These trips, shown as the orange cluster in Figure 8, are sharply separated from all other clusters, highlighting a clear modal distinction.

The urban trip-chain cluster (green in Figure 8) consists of trips with long durations and travel distances but relatively short origin–destination straight-line distances. These trips also exhibit high shares of weekday and peak-hour travel, suggesting multi-stop, activity-based chains such as home–school–work or home–errand–work sequences within dense urban environments. In contrast, the suburban trip-chain cluster features trips with longer distances and higher speeds but similarly short straight-line distances, consistent with chained errands or dispersed activities occurring outside peak periods, often in auto-oriented suburban settings.

Finally, the long-driving cluster captures trips that are substantially faster and longer than the initially labeled driving trips reported in Table 2. These trips likely reflect freeway-

[4] Reported distances and speeds may exceed true values due to spatial uncertainty inherent in mobile phone location data.

oriented travel, whereas the initially labeled driving trips are predominantly local in nature. In summary, these clusters reveal meaningful heterogeneity within the "other" category and provide a principled basis for consolidating trip types into final mode labels in the subsequent analysis.

Table 3: Conditional feature means for different clusters in other trips.

Mode	No. of trips	$\bar{t}$ (h)	$\bar{d}$ (km)	$\bar{l}$ (km)	$\bar{v}$ (km/h)	$\bar{k}$	$\bar{w}$
Active modes	15,241	0.56	6.5	0.75	11.7	0.37	0.67
Urban trip chain	8,317	1.28	23.3	3.46	20.0	0.42	0.75
Suburban trip chain	4,192	0.34	15.8	3.96	48.6	0.26	0.69
Long driving	7,301	0.83	29.1	19.71	38.5	0.41	0.69

By combining the preliminary mode labels with the cluster-based classifications of the "other'' trips, we assign each proper trip a refined and detailed mode label. Figure 9 presents scatter plots of all proper trips in terms of trip speed, trip duration, and straight-line distance, with colors indicating the inferred detailed mode categories. The separation across modes is visually pronounced: each cluster occupies a distinct region of the feature space, underscoring the behavioral differences captured by the inference framework.

In particular, trips classified as active modes are concentrated in the low-speed, short-distance region, while long-driving trips populate the high-speed, long-distance tail. Urban and suburban trip chains occupy intermediate but clearly distinguishable regions, reflecting their contrasting combinations of speed, distance, and trip structure. Notably, driving and transit trips tend to lie near the center of the plots but remain well separated from one another. Driving trips are characterized by shorter durations, higher average speeds, and slightly larger straight-line distances, whereas transit trips exhibit longer durations and lower speeds, consistent with access, waiting, and transfer components. These patterns provide intuitive validation of the inferred mode labels and demonstrate that the framework successfully recovers meaningful modal distinctions from sparse location intelligence data.

Although the urban trip chains, suburban trip chains, and long driving trips exhibit trip characteristics that differ from those of the initially labeled direct driving trips, they are unlikely to represent transit or active-mode travel. Accordingly, we consolidate these categories into the final driving class. Specifically, we assign $m_j = c$ if trip j carries a detailed mode label of "driving," "long driving," "urban trip chain," or "suburban trip chain;" we assign $m_j = p$ if trip j is labeled "transit," and $m_j = a$ if trip j is labeled "active modes."

Table 4 reports the resulting mode class sizes along with the conditional means of key trip attributes for the final mode categories: driving, transit, and active modes. We emphasize that the driving category also includes trips in which the traveler is a passenger in a vehicle

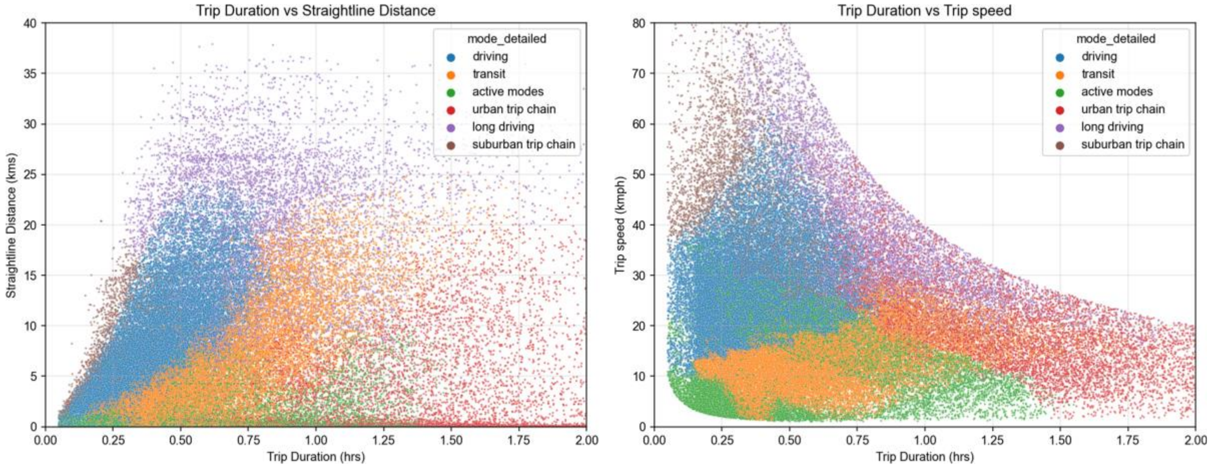

Figure 9: Scatter plots of selected features of all proper trips: straightline distance vs. trip duration and trip speed vs trip duration.

(e.g., family car, taxi, or ride-hailing services such as Uber or Lyft), as it is not possible to reliably distinguish drivers from passengers using mobile phone trajectory data alone. Reassuringly, the resulting mode split among proper trips closely aligns with the benchmark mode shares for Chicago reported in the Chicago Metropolitan Agency for Planning (CMAP) travel survey (Comeaux, 2021), providing an important external validation of our inference approach.

Table 4: Conditional feature means for different modes.

m_j	No. of trips (%)	CMAP %	$\bar{t}$ (h)	$\bar{d}$ (km)	$\bar{l}$ (km)	$\bar{v}$ (km/h)	$\bar{k}$	$\bar{w}$
Driving (c)	49,079 (56.4%)	56.5%	0.60	16.0	8.6	29.8	0.41	0.74
Transit (p)	22,754 (26.1%)	20.5%	0.61	8.4	5.2	12.7	0.44	0.79
Active (a)	15,241 (17.5%)	23.0%	0.56	6.5	0.75	11.7	0.37	0.67

4.3 Transportation insecurity in Detroit

We now turn to training a transportation insecurity (TI) classifier using labeled TSI survey data. The survey data employed in this study come from Wave 18 of the Detroit Metro Area Communities Study (DMACS), conducted between November 2 and December 19, 2023 (Gerber et al., 2025). The DMACS sample includes responses from 2,296 Detroit-area residents; after excluding incomplete records, we retain 2,020 complete responses for analysis. Our primary inputs consist of individual-level socio-demographic variables ($\boldsymbol{y}^{SD}$) , self-reported mode use patterns ($\boldsymbol{y}^{\boldsymbol{m}}$), and responses to the TSI-6 questionnaire, from which we construct the binary transportation insecurity label TI_i based on the sum-of-score method. To enrich the feature set and better capture contextual factors, we further merge census-tract–level living environment variables ($\bar{\boldsymbol{y}}^{LE}$) from the U.S. Environmental

Protection Agency's (EPA) Smart Location Database (U.S. Environmental Protection Agency, 2021), as well as census-tract aggregate socio-demographic variables ($\bar{\boldsymbol{y}}^{SD}$) derived from the American Community Survey (ACS) (U.S. Census Bureau, 2023), using respondents' reported home census tracts as the spatial key.

Table 5 reports descriptive statistics for both the original survey variables and the fused tract-level attributes, providing an overview of the demographic composition, mobility patterns, and neighborhood characteristics represented in the training data.

The survey responses are coded into numerical values to serve as inputs to the TI classifier. The individual-level socio-demographic vector $\boldsymbol{y}_i^{SD}$consists of binary indicator variables with entries in 0*,* 1; consequently, the sample means of these variables correspond to the shares of respondents belonging to each demographic group.

The mode use pattern vector $\boldsymbol{y}_i^{m}$ includes three types of variables. First, mode-specific usage frequencies are converted into probabilities of mode use on a typical day (e.g., "daily" = 1, "once a week" = 0.25, "never" = 0). Second, a car availability indicator is defined, taking the value 0 if the respondent reports "never" using a personal vehicle and 1 otherwise. Third, a multimodality measure is constructed as the count of non-"never" responses across all mode-use questions, capturing the breadth of modes regularly available to the respondent. The census-tract–level aggregate socio-demographic vector $\bar{\boldsymbol{y}}_i^{SD}$ contains the percentages of each demographic group residing in the respondent's home census tract. The living environment vector $\bar{\boldsymbol{y}}_i^{LE}$includes scaled measures of key built-environment attributes, including road density, intersection density, average distance to the nearest transit stop, aggregate transit service frequency, and accessibility by car and transit, measured as the number of jobs reachable within 45 minutes.

Table 5: Descriptive statistics of selected attributes - Detroit.

Category	Variable	Mean	Min	Max	Category	Variable	Mean	Min	Max
$\boldsymbol{y}_i^{SD}$	Female	0.73	0	1	$\bar{\boldsymbol{y}}_i^{SD}$	Female	0.53	0.22	0.67
	Latino	0.09	0	1		Latino	0.10	0.00	0.87
	Native	0.02	0	1		Native	0.00	0.00	0.07
	Asian	0.02	0	1		Asian	0.02	0.00	0.58
	Black	0.71	0	1		Black	0.73	0.00	1.00
	Islander	0.00	0	1		Islander	0.00	0.00	0.05
	White	0.20	0	1		White	0.13	0.00	0.68
	Income 10-30k	0.35	0	1		Income 10-30k	0.27	0.02	0.59
	Income 30-50k	0.16	0	1		Income 30-50k	0.20	0.02	0.53
	Income 50-100k	0.27	0	1		Income 50-100k	0.25	0.01	0.47

	Income 100k+	0.04	0	1		Income 100k+	0.16	0.00	0.70
	Walk	0.46	0	1		0 modes	0.01	0	1
	Bike	0.07	0	1		1 mode	0.25	0	1
	Motorcycle	0.01	0	1	$\boldsymbol{y}_i^m$	2 modes	0.28	0	1
	Own car	0.67	0	1		3 modes	0.22	0	1
	Borrow car	0.06	0	1		4 modes	0.15	0	1
$\boldsymbol{y}_i^m$	Passenger	0.17	0	1		Road density	0.25	0.12	0.47
	Taxi	0.11	0	1		Intersect density	1.20	0.28	4.29
	Bus	0.08	0	1	$\overline{\boldsymbol{y}}_i^{LE}$	Dist to transit	3.54	0.51	9.46
	Train	0.02	0	1		Transit freq	7.35	0.00	54.67
	Paratransit	0.02	0	1		Car access	1.64	0.83	4.29
	Car availability	0.87	0	1		Transit access	1.20	0.16	2.43

Because our ultimate objective is to develop a TI classifier that can be transferred to mobile phone–based location intelligence data, we further simplify the input space by aggregating mode-use variables so that only information related to driving, public transit, and active modes is retained. This harmonization ensures consistency between survey-based and trajectory-based feature representations and facilitates downstream transfer learning. The simplified mode use frequency variables are

$$driving = \min\{motorcycle + own\,car + borrow\,car + passenger + taxi, 1\};$$
$$transit = \min\{bus + train + paratransit, 1\};$$
$$active\,modes = \min\{walk + bike, 1\}.$$

To evaluate the effectiveness of simplified mode-use patterns in predicting transportation insecurity, we train and compare three classes of models. The first is a *full model*, denoted $f_{TI}^{full}(\cdot)$, which uses the complete set of inputs, including individual-level socio-demographics $\boldsymbol{y}_i^{SD}$, detailed mode-use patterns $\boldsymbol{y}_i^m$, and census-tract–level living environment variables $\overline{\boldsymbol{y}}_i^{LE}$. The second is a *simplified model*, $f_{TI}(\cdot)$, which retains $\boldsymbol{y}_i^{SD}$ and $\overline{\boldsymbol{y}}_i^{LE}$ but replaces the detailed mode-use vector with a reduced version that includes only mode-use frequencies for driving, public transit, and active modes, along with the corresponding multimodality indicators. The third is an ablated model, $f_{TI}^{ablation}(\cdot)$, which excludes mode-use information entirely and relies only on $\boldsymbol{y}_i^{SD}$ and $\overline{\boldsymbol{y}}_i^{LE}$.

For each model specification, the data are first split into training and test sets. We then conduct k-fold cross-validation on the training set to compare two supervised learning algorithms—XGBoost and random forest—and select the better-performing algorithm for that model. The selected algorithm is subsequently applied to the test set to generate predictions$\widehat{TI}_i$, from which accuracy $\widehat{Acc}$ and recall $\widehat{Rec}$ are computed.

As discussed in Section 3.3, individual-level socio-demographic variables $\boldsymbol{y}_i^{SD}$ are unavailable in location intelligence data. To assess model robustness under this constraint, we repeat the test-set evaluation by replacing $\boldsymbol{y}_i^{SD}$ with census-tract–level proxies $\overline{\boldsymbol{y}}_i^{SD}$, yielding alternative predictions $\overline{TI}_i$and corresponding performance metrics $\overline{Acc}$ and $\overline{Rec}$ for each model.

Table 6: Performance metrics of different models.

Model	$\widehat{Acc}$	$\widehat{Rec}$	Acc	Rec
Full model $f_{TI}^{full}(\cdot)$	0.80	0.72	0.79	0.75
Simplified model $f_{TI}(\cdot)$	0.74	0.67	0.71	0.69
Ablated model $f_{TI}^{ablated}(\cdot)$	0.69	0.71	0.44	0.90

The k-fold cross-validation results consistently favor the random forest algorithm over XGBoost across all three model specifications. Accordingly, random forest is adopted as the final learning algorithm $f_{TI}^{full}(\cdot)$,$f_{TI}(\cdot)$, and $f_{TI}^{ablated}(\cdot)$. Table 6 summarizes the out-of-sample performance metrics for the three models.

As expected, the full model, $f_{TI}^{full}(\cdot)$, outperforms the simplified model $f_{TI}(\cdot)$ across all metrics, with gains of approximately 5–8 percentage points. This performance gap reflects the value of richer mode-use information in predicting transportation insecurity. Figure 10 presents the results of SHapley Additive exPlanations (SHAP) analyses for both models, highlighting the top ten predictive features and their marginal effects on the probability of experiencing transportation insecurity.

The SHAP results reveal that several detailed mode-use variables—such as the frequency of driving one's own car, riding as a passenger, and using taxis or ride-hailing services—play a critical role in TI classification. Importantly, these modes exhibit opposing effects: frequent use of one's own car is strongly associated with transportation security (high values correspond to a lower predicted probability of TI), whereas reliance on being a passenger or on taxis is positively associated with transportation insecurity. Because the simplified model aggregates all driving-related behaviors into a single category, it is unable to distinguish between these qualitatively different forms of car use, which explains part of its reduced predictive power.

Nonetheless, the simplified model captures many of the key structural predictors of TI. In both the full and simplified models, higher income and reliance on a single mode are negatively associated with TI, while frequent transit use and female gender are positively associated with TI. Figure 10 further shows that overall driving frequency remains a signal of transportation security in the simplified model, although its predictive strength is weaker than that of income and transit use. Taken together, these results suggest that while detailed mode-use patterns improve classification performance, aggregated mode indicators still retain substantial explanatory power and are suitable for transfer to location intelligence data. Overall, when the simplified model $f_{TI}(\cdot)$ is applied using census-tract-level socio-demographic and living environment attributes together with individual-level simplified mode use patterns, it correctly predicts a respondent's transportation insecurity status 71% of the time and identifies 69% of respondents experiencing TI. In contrast, the ablated model—relying solely on census-tract aggregate inputs—fails to achieve a reasonable balance between accuracy and recall. It systematically over-predicts transportation insecurity, leading to substantially poorer classification performance. These results highlight the critical role of individual-level mobility behavior in identifying transportation insecurity: even highly aggregated mode use patterns provide essential information that cannot be recovered from neighborhood-level characteristics alone.

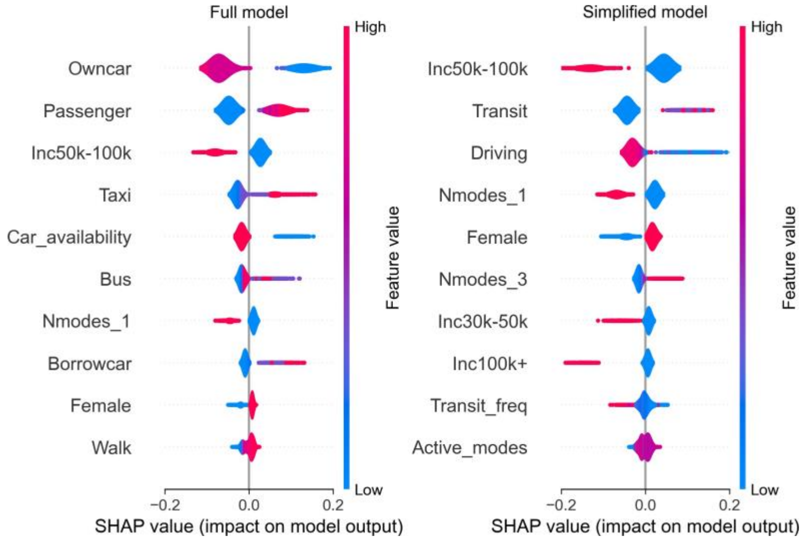

Figure 10: SHAP summary plots of the full model and the simplified model.

4.4 Transportation insecurity in Chicago

Finally, we apply the trained TI classifier $f_{TI}(\cdot)$ to infer transportation insecurity patterns among Chicago residents. This requires transforming the trip-level dataset into device-level mode use patterns, $\boldsymbol{y}_i^m$,which are then fused with census-tract–level living environment $\overline{\boldsymbol{y}}_i^{LE}$ from the EPA Smart Location Database and aggregate socio-demographic variables $\overline{\boldsymbol{y}}_i^{SD}$ from the American Community Survey.

Mode use frequency variables for driving, transit, and active modes are constructed by counting the number of days on which a given mode is observed for each device. These counts are then normalized by the total number of days for which the device has valid records, yielding an empirical probability of daily mode use. This approach ensures consistency with the simplified mode use representation employed in training the TI classifier.

Table 7 reports the descriptive statistics of the resulting Chicago study-device sample. Relative to the Detroit survey respondents, Chicago residents in our sample exhibit higher average incomes, greater reliance on public transit, improved transit service levels, and greater racial diversity. These contextual differences provide a meaningful test of the transferability of the proposed framework across cities with distinct mobility and socio-economic profiles.

Owing to structural differences between Detroit and Chicago, a direct, unadjusted application of the classifier $f_{TI}(\cdot)$ would likely yield biased TI estimates. Indeed, a naive application classifies approximately 75% of the Chicago study devices as transportation insecure—substantially higher than the estimated share in Detroit (36%) (Wileden and Murphy, 2025) and the national average (17%) (Murphy et al., 2024). This discrepancy is not surprising, given differences between the two cities in income distributions, transit supply, travel behavior, and urban form.

Table 7: Descriptive statistics of selected attributes - Chicago.

Category	Variable	Mean	Min	Max	Category	Variable	Mean	Min	Max
$\boldsymbol{y}_i^m$	Driving	0.37	0.00	1.00	$\bar{\boldsymbol{y}}_i^{SD}$	Female	0.52	0.08	0.67
	Transit	0.23	0.00	0.86		Latino	0.30	0.00	0.99
	Active modes	0.15	0.00	0.97		Native	0.01	0.00	0.16
	1 mode	0.15	0	1		Asian	0.06	0.00	0.84
	2 modes	0.21	0	1		Black	0.40	0.00	1.00
	3 modes	0.65	0	1		Islander	0.00	0.00	0.05
$\bar{\boldsymbol{y}}_i^{LE}$	Road density	0.29	0.09	0.69		White	0.29	0.00	0.91
	Intersect density	1.62	0.22	8.15		Income 10-30k	0.18	0.01	0.67
	Dist to transit	2.86	0.00	9.32		Income 30-50k	0.15	0.01	0.42
	Transit freq	31.6	0.0	393.7		Income 50-100k	0.27	0.03	0.56
	Car access	4.04	1.74	7.15		Income 100k+	0.32	0.00	0.83
	Transit access	5.07	0.12	14.80					

While the absence of TSI survey data in Chicago prevents us from recalibrating the classifier or directly validating the inferred TI labels, we can nonetheless present a set of preliminary results that speak to the internal consistency, spatial plausibility, and behavioral relevance of the predictions. These results help illustrate the potential of the proposed transfer learning framework and provide justification for its further refinement and application in future studies.

Assuming that the share of transportation-insecure individuals in the study population is comparable to the national average of 17%, we recalibrate the decision threshold of the classifier to better align predicted prevalence with this benchmark. Specifically, instead of using the default probability threshold of 0.50, we increase the threshold so that a device is classified as transportation insecure only if its predicted probability exceeds 0.645. Under this adjusted threshold, 17.3% of study devices are classified as transportation insecure, closely matching the assumed national reference level. This recalibration allows us to assign a binary TI label (secure or insecure) to every study device while maintaining a plausible overall performance.

Figure 11 maps the resulting share of transportation-insecure devices across census tracts, computed as the proportion of insecure devices among all study devices residing in each tract. The spatial pattern reveals clear clustering, with higher inferred TI shares concentrated in parts of the west and south sides of Chicago. These areas also correspond to neighborhoods that differ systematically in socioeconomic conditions and travel environments, suggesting that the inferred TI patterns are spatially coherent and consistent with known variations in urban mobility conditions.

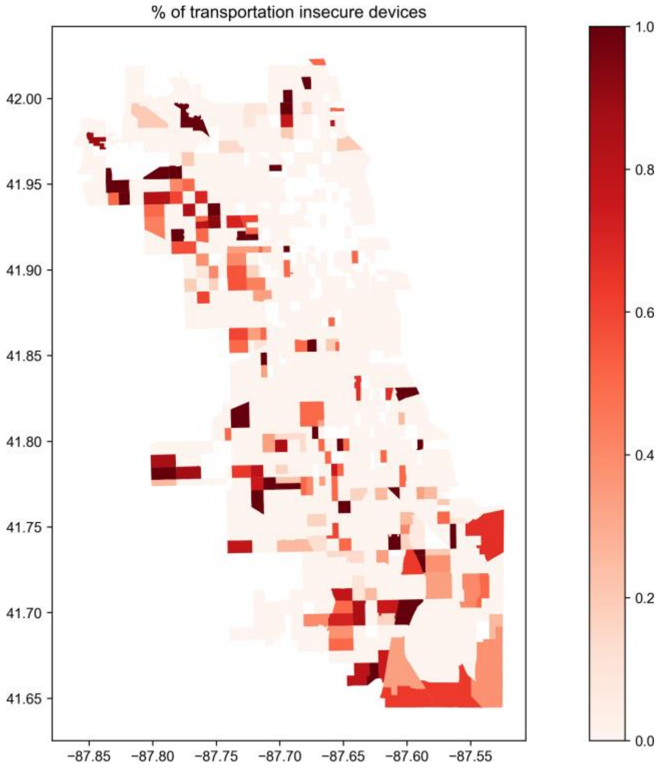

Figure 11: Percentage of transportation insecure devices in each census tract.

5 Conclusion

This project proposed and demonstrated an alternative approach to measuring transportation insecurity (TI) using large-scale location intelligence data. Motivated by growing evidence that TI is closely linked to travelers' mode use patterns, we developed an integrative framework that combines mobility inference, supervised learning, and transfer learning. The framework first infers trip-level travel modes from sparse and irregular mobile phone trajectories using a semi-supervised learning strategy. These trip-level labels are then aggregated to construct device-level mode use patterns, which serve as key inputs to a TI classifier trained on Transportation Security Index (TSI) survey data augmented with census-tract-level con-textual information. Finally, the trained classifier is transferred to location intelligence data after harmonizing feature definitions and scales across data sources.

Empirically, we find that even simplified mode use patterns—capturing reliance on driving, transit, and active modes—carry substantial predictive power for TI. When combined with census-tract-level socio-demographic and living-environment variables, these patterns enable TI classification with 71% accuracy and 69% recall, even in the absence of individual-level socio-demographic data. In contrast, models that rely exclusively on aggregate contextual variables perform poorly, underscoring the critical role of individual mobility behavior in identifying transportation insecurity. These results suggest that mode use patterns provide a meaningful behavioral signal that complements traditional socio-economic indicators.

At the same time, our case study highlights important challenges in transferring TI classifiers across regions. Because the classifier is trained using Detroit survey data, direct application to Chicago reveals sensitivity to structural differences in urban form, transportation systems, and socio-economic conditions. While this limits the immediate accuracy of out-of-context predictions, our preliminary findings still reveal spatially coherent TI patterns in Chicago that align with well-documented disparities in mobility conditions. This outcome illustrates both the promise and the current limitations of transfer learning for TI measurement.

The results from the present study point to several directions for advancing this framework toward operational maturity. First, validation studies are needed in regions where both TSI survey data and location intelligence data are available, allowing direct comparison between survey-based and inferred TI measures. Such studies would provide critical evidence on the reliability and robustness of location-based TI inference. Second, future work should focus on developing a more generalizable TI classifier by integrating information from multiple local contexts. This may involve identifying contextual features that capture structural differences across regions and incorporating domain-adaptation or hierarchical modeling techniques to account for these differences explicitly. Finally, improvements in location intelligence data coverage and representativeness—both in terms of sample size and population coverage—will be essential for scaling the approach and supporting long-term monitoring of TI.

In summary, this project demonstrates that location intelligence data, when combined with principled machine learning and transfer learning methods, can serve as a viable foundation for measuring transportation insecurity at scale. While additional work is needed to address validation and generalizability, the proposed framework opens a promising pathway toward more timely, spatially detailed, and behaviorally grounded assessments of transportation insecurity to support transportation planning and policy evaluation.

6 Acknowledgments

This research is funded by Center for Connected and Automated Transportation (CCAT), part of the University Transportation Center Program sponsored by US Department of Transportation (USDOT).

References

Ahas, R., Silm, S., Järv, O., Saluveer, E., and Tiru, M. (2010). Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of urban technology*, 17(1):3–27.

Alexander, L., Jiang, S., Murga, M., and González, M. C. (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation research part c: emerging technologies*, 58:240–250.

Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, M., and Puchinger, J. (2019). Inferring dynamic origin-destination flows by transport mode using mobile phone data. *Transportation Research Part C: Emerging Technologies*, 101:254–275.

Bader, M., McComas, M., Williams, A., Iyer, S., Locke, D., Parker, A., and Truiett-Theodorson, R. (2025). A portrait of baltimore 2024. Technical report, Johns Hopkins University.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira Jr, J., and Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation research part C: emerging technologies*, 26:301–313.

Chang, E., Zhang, M., Li, Z., and Hu, Y. (2025). When digital disadvantage meets transport disadvantage: The association between smartphone-based mobility services and perceived transport disadvantage among the elderly. *Journal of Transport Geography*, 123:104119.

Chen, C., Bian, L., and Ma, J. (2014). From traces to trajectories: How well can we guess activity locations from mobile phone traces? *Transportation Research Part C: Emerging Technologies*, 46:326–337.

Chen, X., Wan, X., Li, Q., Ding, F., McCarthy, C., Cheng, Y., and Ran, B. (2019). Trip-chain-based travel-mode-shares-driven framework using cellular signaling data and web-based mapping service data. *Transportation Research Record*, 2673(3):51–64.

City of Chicago (2024). Census tracts. Socrata dataset identifier: 4hp8-2i8z.

Comeaux, D. (2021). A pre-pandemic snapshot of travel in northeastern illinois key findings. Technical report, Chicago Metropolitan Agency for Planning.

Dai, T., Bella, G., Kivestu, P., Chen, Y., Stathopoulos, A., and Nie, Y. M. (2025). What mobile phone data reveal about mobility patterns of teleworkers. *Transportation Research Part A: Policy and Practice*, 201:104670.

Fu, X., Zhang, Y., Ortúzar, J. d. D., and Lü, G. (2025). Activity-travel pattern inference based on multi-source big data. *Transport Reviews*, 45(1):26–48.

Gerber, E., Morenoff, J., and Ostfeld, M. (2025). Detroit metro area communities study (dmacs) wave 18, michigan, 2023.

Gould-Werth, A., Griffin, J., and Murphy, A. K. (2018). Developing a new measure of transportation insecurity: an exploratory factor analysis. *Survey Practice*, 11(2).

Grengs, J. (2012). Equity and the social distribution of job accessibility in detroit. *Environment and Planning B: Planning and Design*, 39(5):785–800.

Hariharan, R. and Toyama, K. (2004). Project lachesis: parsing and modeling location histories. In *International Conference on Geographic Information Science*, pages 106–124. Springer.

Huang, H., Cheng, Y., and Weibel, R. (2019). Transport mode detection based on mobile phone network data: A systematic review. *Transportation Research Part C: Emerging Technologies*, 101:297–312.

Lind, E. M., Owen, A., Liu, S. S., and Hockert, M. (2025). Accountability through accessibility: Measuring what matters for departments of transportation. Technical report, Center for Transportation Studies, University of Minnesota.

Liu, Y., Liao, F., Wang, W., Wang, Y., and Chen, J. (2025). An integrated method for inferring multimodal travel mode choices using mobile network data. *Transportation Research Part C: Emerging Technologies*, 179:105305.

McDonald-Lopez, K., Murphy, A. K., Gould-Werth, A., Griffin, J., Bader, M. D., and Kovski, N. (2023). A driver in health outcomes: developing discrete categories of transportation insecurity. *American journal of epidemiology*, 192(11):1854–1863.

Murphy, A. K., Gould-Werth, A., and Griffin, J. (2021). Validating the sixteen-item transportation security index in a nationally representative sample: A confirmatory factor analysis. *Survey Practice*, 14(1).

Murphy, A. K., Gould-Werth, A., and Griffin, J. (2024). Using a split-ballot design to validate an abbreviated categorical measurement scale: An illustration using the transportation security index. *Survey Practice*, 17:1–41.

Murphy, A. K., McDonald-Lopez, K., Pilkauskas, N., and Gould-Werth, A. (2022). Transportation insecurity in the united states: a descriptive portrait. *Socius*, 8:23780231221121060.

Murphy, A. K., Pilkauskas, N. V., Kovski, N., and Gould-Werth, A. (2025). How does transportation insecurity compare and relate to other indicators of material hardship in the us? *Social Indicators Research*, pages 1–29.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Pappalardo, L., Ferres, L., Sacasa, M., Cattuto, C., and Bravo, L. (2021). Evaluation of home detection algorithms on mobile phone data using individual-level ground truth. *EPJ data science*, 10(1):29.

Park, J., Gorecki, M., Flicker, A., Reyner, A., Henize, A., Klein, M., and Beck, A. F. (2025). Illustrating the role of infant well child visits in the association between transportation insecurity and acute healthcare use. *Academic Pediatrics*, page 102827.

Phillips, A., Coughenour, C., and McDonough, I. (2025). Transportation insecurity in older adults aged 60 and older in clark county, nevada. *Journal of Transport & Health*, 45:102196.

Pohl, A. L., Aderonmu, A. A., Grab, J. D., Cohen-Tigor, L. A., and Morris, A. M. (2025). Transportation insecurity, social support, and adherence to cancer screening. *JAMA Network Open*, 8(1):e2457336–e2457336.

Robbennolt, D., Beliveau, A., and Bhat, C. R. (2025). An investigation of physical participation dissonance and virtual activity participation in the united states. *Transportation Research Part A: Policy and Practice*, 201:104696.

Singer, M. E. and Martens, K. (2023). Measuring travel problems: Testing a novel survey tool in a natural experiment. *Transportation research part D: transport and environment*, 121:103834.

Smart, M. J., Klein, N. J., Consortium, M. N. T. R., et al. (2015). A longitudinal analysis of cars, transit, and employment outcomes. Technical report, Mineta National Transit Research Consortium.

U.S. Census Bureau (2023). American community survey 1-year estimates. Table DP03 and DP05.

U.S. Environmental Protection Agency (2021). Smart location database (version 3.0).

Wander, M., Blumenberg, E., Brozen, M., and Butler, T. L. (2025). The promise of universal basic mobility. *Transport Reviews*, pages 1–21.

Wileden, L. and Murphy, A. K. (2025). Transportation insecurity in the motor city. Technical report, University of Michigan.

Zhong, S., Chen, J., and Cai, M. (2024). A transport mode detection framework based on mobile phone signaling data combined with bus gps data. *Mathematics (2227-7390)*, 12(23).

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.

A Final documentation of output, outcomes and impacts

A.1 Outputs

1. A transfer learning framework that allows for identifying transportation insecure population using large-scale location intelligence data.
2. A paper entitled “What mobile phone data reveal about mobility patterns of teleworkers,” presented at various conferences (2024 INFORMS, Seattle, WA; 104th TRB Annual Meeting, Washington, D.C.) and published in the academic journal *Transportation Research Part A: Policy and Practice*.
3. A presentation entitled “Learning transportation insecurity from mobile phone data,” at 2025 CCAT Global Symposium on Mobility Innovation, Ann Arbor, MI.

A.2 Outcomes

1. Developed a suitable data processing and analysis framework for activity, trip and mobility features extraction.
2. Proved validity of classifying travelers based on mobility features.
3. Understood features associated with transportation insecurity.
4. Developed a quantitative supervised machine learning model that predicts transportation insecurity status based on various predictors.
5. Understood features related to transit trips, driving trips and other trips.
6. Experimented a few machine learning methods in inferring travel modes.
7. Two PhD students and one undergraduate student received training through the project.

A.3 Impacts

1. The findings enhance our understanding of transportation insecurity, which may help policy makers find solutions to the problem.
2. The project contributes to workforce development by exposing graduate and undergraduate students to research in data mining and transportation modeling.

3. The project reveals more details about commercial mobile phone data and their potentials in trip-level information inferences

4. The project suggests potential usage of mobile phone data in guiding transportation policy making.

A.4 Collaborating organizations

1. Oplytix, LLC. Minneapolis, MN. Provided data cleaning, storage and preprocessing services.

2. Gerald R. Ford School of Public Policy, University of Michigan. Ann Arbor, MI. Provided engineering and technical support on data sources.

A.5 Challenges/Problems

1. Due to unavailability of location intelligence data and survey data in the same city, validation results were preliminary.

2. Project end date changed to 1/31/2026.