Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2023) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

25th Euro Working Group on Transportation Meeting (EWGT 2023)

GPS-Based Trip Phase and Waiting Time Detection to and from Public Transport Stops Via Machine Learning Models

Seyed Hassan Hosseini[a]*, Siavash Pourkhosro[a], Guido Gentile[a], Lory Michelle Bresciani Miristice[a]

[a]*Sapienza University of Rome, Via Eudossiana, 18, 00184 Rome, Italy*

Abstract

Recognizing passengers' trip phases can provide valuable insights for urban planners in making decisions regarding urban planning, as it involves identifying the various stages of a passenger's journey. There is a significant research gap in current trip phase recognition research. Most existing works rely on traditional techniques such as GIS, surveys, and direct observation at stations to handle the task. The goal of this paper is to present a novel approach to determine the time and distance of access and egress trip phases and waiting time at bus stations, using a machine learning algorithm based on raw GPS trajectories. Specifically, we train a random forest model using two large datasets, Geolife GPS trajectory, and Sussex-Huawei Locomotion dataset, to detect the various transportation modes. Furthermore, a new algorithm is developed for detecting access time/distance, egress time/distance, and waiting time. Our approach is the first investigation in trip phase recognition that combines two large datasets and a machine learning model for trip phase detection. Our study yields the following accuracies on test trips saved in Roma: access time and distance predicted with 80.68% and 91.61%, while egress time and distance arrived at 72.63% and 70.68% accuracy. Passenger waiting time predicted from raw GPS data as a new feature with 82.01% accuracy at the bus station. These results underscore the effectiveness of our approach in predicting different phases.

Keywords: Trip Phase Recognition; Public Transportation; Machine Learning; GPS trajectories, Access/Egress Phase

1. Introduction

Passenger Travel patterns show significant variations in time and location due to urban expansion and functional division. Individuals in large cities often use a variety of modes for daily travel, and understanding urban trip phases in public transport based trips can be essential to enhance urban planning. Such understanding enables us to identify

* Corresponding author. Tel.: +39- 3515765995
E-mail address: seyedhassan.hosseini@uniroma1.it

Peer-review under responsibility of the scientific committee of the 25th Euro Working Group on Transportation Meeting (EWGT 2023)

areas of concern and devise practical solutions to address them. Mobile phones, as a capturing tool to record passenger movements via GPS and inertial sensors, offer a viable option for recognizing the behavioral patterns of individuals within urban areas. However, traditional methods like surveys and telephone and email interviews can be costly and time-consuming. Automating trip phase recognition is the key objective of this study. Passenger data collection and filtering are among the first steps. Our ultimate goal is to extract features from GPS points, train a machine learning model, and finally detect distinct journey stages.

2. Literature Review

This section involves a detailed examination of two primary stages: the initial step involves determining the mode of transportation, whereas the second step explains the recognition of the passenger's trip phase.

2.1. Transport Mode Detection

The first step in trip phase identification is transport mode recognition for each chuck of GPS data and each study tries to follow a specific approach and use different sensor data. To categorize mobile sensor data in this study Feng et al. (2016) several techniques such as Naive Bayesian, Bayesian network, logistic regression, multilayer perceptron, and support vector machine were applied with different final accuracies to categorize data into several modes of transport. Moreover, GPS data as a primary source of passenger data was used by Stenneth and Wolfson (2011) to identify various modes of transit. In this study the performance of five distinct inference models, namely Bayesian Net, Decision Tree, Random Forest, Naive Bayesian, and Multilayer Perception, in identifying transport modes such as car, bus, train, walking, cycling, and standing and random Forest algorithm were evaluated and arrived to superior performance, achieving an accuracy of 93.7%. The study conducted by Stenneth and Wolfson (2011) had two limitations. The lack of trip segment identification was the first issue, and a small training dataset with just six participants was the second drawback Sadeghian (2021).

In another study Stopher (2008), a probability matrix was used to find out five transportation modes including walk, bike, car, bus, and train based on data from GPS devices. They considered four motion characteristics such as average speed, maximum speed, most often recorded speed, and distance to define the probability of a specific transport mode and after that, they added a GIS layer of information of the transport network to differentiate between bus and car. Therefore, the primary variables in this study are speed and GIS network and their statistical method achieved an accuracy of 95%.

The integration of GIS and other environmental passenger data has the potential to enhance the final accuracy of transport mode detection models. A study conducted by Biljecki (2013), used a fuzzy logic tries to detect transportation modes from GPS trajectories. The accuracy of their study was rather low when they did not use an additional GIS layer. The results showed that the accuracy of the model became 92% after adding the GIS network. Fuzzy logic rules have one limitation as compared to machine learning techniques as depend on numerous defined rules that depend on a particular scenario. Using fuzzy logic rules requires more human involvement than using machine learning techniques Sadeghian (2021).

The goal of this paper Xiao (2015) was to classify various modes into walking, bike, e-bike, bus, and car from GPS data using four machine learning algorithms (SVM, MNL, ANNs, and BN) and showed that the BN algorithm performed more accurately than other algorithms (92%).

Deep learning algorithms have demonstrated their capability to classify various modes of transportation. In a relatively recent study Yazdizadeh (2019) a CNN algorithm based on GPS dataset to detect four modes of transport (walking, bike, car, and public transport) was developed. They combined CNN outputs with majority voting, average voting, and optimal weight techniques. Moreover, a random forest model as a meta-learner was trained to use an ensemble library and reached an accuracy equal to 91.8%. This paper Dabiri (2018) used the Geolife dataset to predict five transport modes: walking, bike, car, bus, and train. They tested two unsupervised and five supervised algorithms, including KNN, SVM, DT, RT, and MLP, and found that the CNN algorithm outperforms other methods. The results indicated accuracy equal to 85 percent. Similarly, Yu (2020) employed the same dataset as Dabiri (2018). However, they could increase mode detection accuracy by using a semi-supervised deep ensemble learning method.

2.2. Trip Phase Recognition

The second step is the recognition of trip phases via the results of the transport mode detection. One approach to identifying trip phases involves traditional methods, such as questionnaires, and direct observations at public transit stations. However, these conventional methods are time-consuming, costly, and susceptible to human errors. There is a research gap in this domain, prompting our proposal for a novel approach that leverages raw GPS data and machine learning algorithms to overcome limitations. This paper Hosseini and Gentile (2022) used accelerometer data from smartphones to recognize different trip phases, focusing on identifying the access phase and waiting time at public transport stops. In our study, raw GPS data from combining two datasets, Geolife and Sussex, are the primary source for training our machine learning model. In addition, GPS trip data was recorded in Rome to evaluate model effectiveness and validation purposes.

Direct observation method Kim (2015) was applied to identify pedestrian walking distance, and 139 individuals were chosen from two transit stations and followed up to their final destinations, revealing an average walking distance of 548 meters. In another study, Tennøy and Knapskog (2022) questionnaires were used to identify the durations and distances of walking trips to public transport in small Norwegian cities. In this paper, Jinliao and Ruozhu (2018), a questionnaire survey and face-to-face interview were applied to examine the connection between demographic characteristics and walking access distance at Nanjing metro line 2 in China.

Two distinct methods of data collection Nygaard (2016) were employed to compare the actual waiting times of passengers with those provided by transport models. First, the exact waiting time of 1145 passengers at bus stations was recorded. Second, they surveyed 109 public transit passengers at stations to gather information on their waiting time strategies, which pertain to how they plan their arrival time at public transport stops. Another investigation El-Geneidy et al. (2014) detailed origin-destination information extracted from users to generate service areas that define walking catchment areas around transit services in Montreal, Canada. The authors employed the 85th percentile walking distance to bus transit service for home-based trip origins and home-based commuter rail trip origins, which were found to be 524 meters and 1259 meters, respectively, based on the survey information.

GPS-based Household Travel Survey data from 2009-2010 in the Cincinnati metropolitan region was used as a main source in this investigation Ting and Zuo (2018) to delineate the transit catchment area. Researchers employed GIS techniques to estimate the distance between transit stops and the corresponding walking/cycling routes. Specifically, they projected GPS points onto the road network to determine the most suitable travel methods. True walking distance rather than straight-line distance was the primary purpose of this research Alan Hoback (2008). They employed a GIS network and a Monte Carlo simulation to find the walking distance to and from a bus stop. Aside from all prior investigations, our proposed technique is expected to provide accurate and efficient trip phase detection, which enhances the overall operation of public transportation systems by offering insights into travel patterns, passenger behavior, and areas for improvement.

The paper is organized as follows: section 3 discusses the methodology; the validation procedure and results are covered in section 4. Section 5 presents conclusions.

3. Methodology

Data preprocessing is paramount in this study, as unprocessed GPS data contain errors and outliers. In the following sections, we present an in-depth clarification of the datasets, demonstrate the methodology employed for data cleansing, and evaluate the techniques employed for trip identification and segmentation.

3.1. Datasets

The current investigation entails combining two large datasets. Geolife project by Microsoft Research Asia, recorded by Zheng (2011) and GPS points gathered from 182 participants from April 2007 to October 2011. It contains log trajectories with a 1-5 second frequency. Furthermore, 69 users labeled transportation modes throughout their trips. The second dataset H.Gjoreski and M.Ciliberto (2018) , a comprehensive collection of smartphone sensor data for transportation analysis. This dataset was captured over seven months in 2017 from three

participants who recorded eight unique modes of transportation. Standing labels were extracted from Sussex-Huawei data to use in our analysis.

Furthermore, new GPS trajectories recorded via a GPS logger Android application, including twenty trips across Rome, with one second frequency.

3.2. Data Cleaning

The initial step in the preprocessing phase involved the removal of all Geolife data outside the timeframe of April 2007 to October 2011. Furthermore, Geolife and Sussex-Huawei datasets consist of geographical coordinates, especially latitude, and longitude, which could potentially contain errors. To address this, we removed any latitude and longitude values that exceed the ranges of -90 to 90 and -180 to 180, respectively.

During the following stage, we checked whether two consecutive timestamps in each journey are the same. If we found such an occurrence, we removed one of the timestamps. Moreover, in accordance with research carried out by Dabiri (2018), we established speed and acceleration limits for walking, cycling, bus, and train (Table 1). Points exceeding these predefined limitations in terms of speed or acceleration are subsequently eliminated. Moreover, we assumed a maximum speed limit for standing mode. A minimum speed for cycling is considered based on the average value of all the minimum bike speeds that Kassim (2020) suggested in their research. The last presumption was to consider walking segments in Geolife dataset with an average speed of less than 0.99 m/s as standing.

Table 1. Speed and Acceleration Limitation

Transportation Mode	Maximum Speed (m/s)	Maximum Acceleration (m/s^2)
Walk	7	3
Bike	12	3
Bus	34	2
Train	34	3
Standing Still	1	-

3.3. Trip identification

Since datasets are made from a collection of timestamp points and different trips are not explicitly distinguished from one another, trips are categorized on specific criteria; when the time interval between two consecutive GPS points exceeds 20 minutes, then the journey is finished, and a new trip id is assigned for the following GPS points Zheng (2011) and Dabiri (2018). Finally, a total number of 5770 trips were identified to train our machine learning model, and 153 out of the 5,770 trips were selected for the trip phase recognition, assuming they started and ended with walking or cycling mode.

3.4. Trip Segmentation

A two-step algorithm was applied to examine unique trajectories to achieve equal segments. First, trips are segmented into various segments according to the recorded modes of transportation. After that, trip chunks obtained from the first step are divided into smaller segments, which include 100 GPS points. Finally, segments of equal size are acquired, comprising 100 GPS points, and linked to separated modes of transportation.

3.5. Motion Features

Different motion characteristics are calculated for each segment. The Vincenty formula was initially applied to calculate the distance between two points on an ellipsoidal earth model. The distance between each successive GPS point was calculated on each trip. In addition, we compute the speed and speed variations between two GPS locations as acceleration and the variations in acceleration between two GPS points as jerks. Another motion

characteristic is bearing rate, which represents the velocity of change in orientation across different transportation modes. Pedestrians and cyclists demonstrate higher maneuverability rates than buses and trains due to their ability to change their path while in motion. Conversely, buses and trains are forced to follow their assigned routes.

An analysis applied to motion features to identify distinct characteristics for each segment. The initial metric involves the mean velocity of every segment, computed by dividing the aggregate distance of each 100 GPS point by the total time duration of the segment. The second one is expectation speed which is the average of all the speeds in the segment. The low-speed ratio is another feature equal to the number of GPS points with a speed less than a predefined threshold divided by the total number of GPS points in each segment, which is 100 for each segment. Moreover, the three largest speeds, accelerations, standard deviation of bearing rate, and average jerk for each segment were chosen as new features.

3.6 Transport Mode Detection

The initial stage involves the prediction of transportation modes for every set of 100 GPS points. After combining two large datasets and training a random forest model with motion features (speed, jerk, etc.), the model receives data from each time frame and predicts each mode of transport. Results arrived at an accuracy performance of 93.65 percent. To find the optimum hyper-parameter combination, we investigated a variety of combinations, and fixed number of estimators equal to 150 and fixed the bootstrap as true.

3.7 Trip Phase Recognition

The main objective of this research study is to identify the various stages of a journey taken by passengers, from the start part of a trip to a destination. These stages of travel can include access, motorized, and egress phases. The access phase refers to the initial mode of transportation used to reach the first public transit stop, such as walking or cycling. The motorized phase pertains to the primary mode of transportation to reach the destination, such as taking a train or bus. The egress phase denotes the concluding stage of the trip, where the traveler employs a mode of transportation to reach the ultimate destination from the last transit stop, such as walking or cycling (Fig 1).

In this study, we propose a novel approach for determining the time and distance of the access and egress phases and the waiting time at bus stations using raw GPS trajectories and machine learning algorithms. To achieve this objective, it is essential to have prediction results of transport modes associated with each GPS point, such as standing, walking, cycling, or taking a bus or train. As mentioned in the previous section, we employed a random forest model to predict transportation modalities for every trip, including several chunks (100 GPS points). Subsequently, the GPS trajectories were annotated with transport labels, facilitating the development of algorithms to differentiate between the access, motorized, and egress stages of a trip.

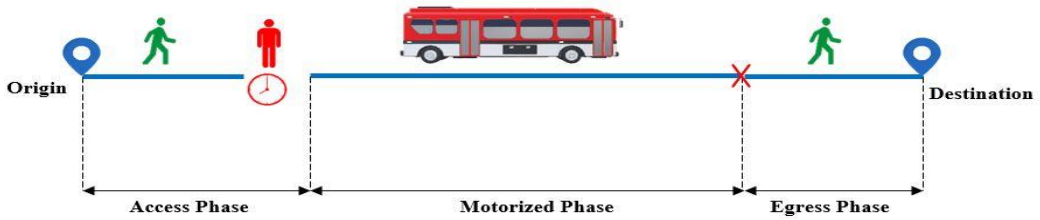

Fig 1. Different Phase of Public Transport Trips

To develop algorithms to detect trip phases and waiting time at bus stations, we selected trips from the Geolife dataset that start with walking or cycling, continue with a bus or train, and end with walking or cycling. Selected trips were not included in the training of the random forest model. The algorithm starts from the beginning of a journey and progresses towards the destination until reaching a motorized section. The motorized phase is assumed to begin when two consecutive bus/train segments are encountered. The segments preceding this motorized phase are considered the access portion of the trip. A similar procedure is followed in reverse to identify the egress phase.

Starting from the destination and moving towards the origin, whenever two successive bus/train segments are encountered, the motorized phase is observed, and the segments after this point are selected as the egress portion.

Continuing the process, the access time for the access phase is computed by summing the delta time between every two consecutive points within this phase. The procedure for determining access distance involves summing the distances between every two points in the access phase, resulting in the computation of the total access distance. The same procedure is applied to the egress phase in a similar manner.

It is worth mentioning that we utilize a supervised machine learning algorithm to predict transport modes. The GPS trajectories used to train this model are already labeled with the corresponding transportation modes, and we refer to these trips as actual trips. On the other hand, after applying random forest model, we have trips with new predicated transportation modes, and these new predicted modes might differ slightly from the actual transportation modes due to the accuracy of the random forest model. We refer to these trips as predicted trips.

a

b

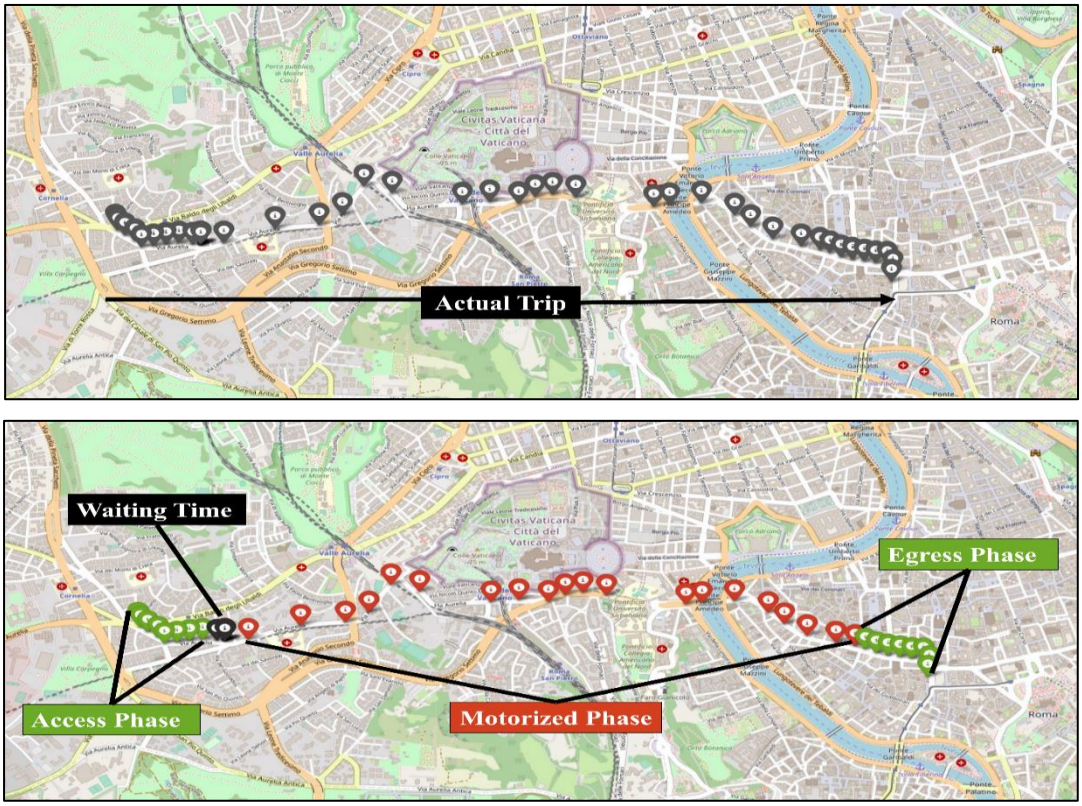

Fig 2. (a) Actual: recorded trip before using our approach which is only raw GPS points; (b) Prediction: after using our approach different phases are specified with different colors

4. Validation and Results

We compute the access, egress time, and distance for actual and predicted trips. Mean Absolute Percentage Error (MAPE) was applied as a primary indicator to calculate the absolute accuracy in predicting each journey phase. The MAPE is the mean of the absolute percentage errors between the predicted and actual values. Applying this method, we obtained accuracies of 81.47% for access time, 82.19% for access distance, 94.74% for egress time, and 93.80% for egress distance.

One of the key contributions of this study was the computation of waiting time at bus stops directly from GPS points. We calculate the waiting time for selected Geolife trips. There is no labeled data for standing in the Geolife dataset, and the solution is extracting standing labeled data from Sussex datasets, enabling us to accurately predict the standing mode (waiting time) in our selected trips. Our algorithm for detecting waiting time begins by identifying the access phase of a trip, following a similar process to the one used for detecting access times. Starting from the beginning of the trip, we halt the algorithm whenever two consecutive bus segments are encountered, marking this as the access phase. Within the access phase, we traverse the segment sequence in reverse order, searching for successive standing segments until we reach a bike or walk segment. Ultimately, the waiting time can be determined by summating the time intervals between two consecutive points across all consecutive stationary segments extracted from our algorithm.

In another scenario, the algorithm is programmed to identify "Standing" segments that occur between a "Walk/bike" segment and a triple segment consisting of a "walk" segment followed by two consecutive "bus" segments. The decision to include a "walk" segment before two consecutive bus segments is made to account for potential misclassifications, such as when a bus starts moving slowly from a station, causing the model to mistakenly classify it as walking. If we do not consider this scenario, the previous algorithm would be unable to detect waiting times for these types of trips, resulting in the loss of computing waiting times for some trips where the model may incorrectly predict the transport mode as walking instead of a bus.

Trips that were not included in the training process stage are fed into our model to calculate the duration of waiting time. All trips for testing phase have been derived from the Geolife dataset to validate the generality aspect of our algorithm. A two-stage classification algorithm applied for each single trip. The first step involves the use of a random forest model to detect the mode of transportation. The second step employs the findings to identify the trip phase of each urban trip. Moreover, our methodology can define the duration of stationary periods (i.e., waiting times) during these journeys.

Finally, we evaluate the effectiveness of our algorithms by testing on a newly collected dataset in Rome. This dataset includes information such as latitude, longitude, and timestamp, and a GPS logger application records trajectories at a frequency of 1 second. Additionally, we capture transportation modes, access/egress time, and waiting time to compare them with the algorithms' predictions. We employ the mean absolute percentage error method to calculate accuracy. The results demonstrate that the algorithms achieve an accuracy of 82.01% for waiting times, 80.68% for access time, 91.61% for access distance, 72.63% for egress time, and 70.68% for egress distance.

Table 2. Results of trip phase recognition algorithm

Trip Phase	GeoLife	Roma dataset
Access Time	81.47	80.68
Access Distance	82.19	91.61
Egress Time	94.74	72.63
Egress Distance	93.80	70.68
Waiting Time	------	82.01

5. Conclusion

The results demonstrate that our algorithm has achieved a significant level of accuracy across various phases, highlighting its potential as an alternative to conventional approaches such as GIS-based surveys and direct observation for computing access and egress time and distance. However, additional research is necessary to enhance the suggested approach. A viable strategy for improving the accuracy of the proposed method is to use deep learning algorithms to predict transportation modes. Moreover, more trips are required to validate the efficacy of this approach. Therefore, collecting additional data sets or utilizing existing ones is recommended. GPS data are insufficient to determine waiting times due to the lack of signals at metro stations. Therefore, an alternative solution

to improve the suggested approach involves using mobile sensor data (accelerometer) to determine waiting times at Metro stations accurately.

References

Alan Hoback, S. A. (2008). True Walking Distance to Transit. *Transportation Planning and Technology* , 681-692.

Biljecki, F. (2013). Transportation mode-based segmentation and classification of movement trajectories. *International Journal of Geographical Information Science*, 385-407.

Dabiri, S. (2018). Inferring transportation modes from GPS trajectories using a convolutional neural network. *Transportation Research Part C: Emerging Technologies*, 360-371.

El-Geneidy, A., & Grimsrud, M. (2014). New evidence on walking distances to transit stops: identifying redundancies and gaps using variable service areas. *Transportation volume 41*, 193–210.

Feng, T. (2016). Comparison of advanced imputation algorithms for detection of transportation mode and activity episode using GPS data. *TRANSPORTATION PLANNING AND TECHNOLOG*.

Gjoreski, H. (2017). *The University of Sussex-Huawei Locomotion and Transportation Dataset* .

Hosseini, S. H., & Gentile, G. (2022). Smartphone-based recognition of access trip phase to public transport stops via machine learning models. *Transport and Telecommunication,Riga*, 273-283.

Jinliao, H., & Ruozhu, Z. (2018). Walking Access Distance of Metro Passengers and Relationship with Demographic Characteristics: A Case Study of Nanjing Metro. *Chinese Geographical Science*, 612-623.

Kassim, A. (2020). Critical review of cyclist speed measuring techniques. *Journal of Traffic and Transportation Engineering*, 98-110.

Kim, H. (2015). Walking distance, route choice, and activities while walking: A record of following pedestrians from transit stations in the San Francisco Bay area. *URBAN DESIGN International*, 144-157.

Nygaard, M. F. (2016). Waiting Time Strategy for Public Transport Passengers. *Annual Transport Conference at Aalborg University.*

Sadeghian, P. (2021). Review and evaluation of methods in transport mode detection based on GPS tracking data. *Journal of Traffic and Transportation Engineering*, 467-482.

Stenneth, L., & Wolfson, O. (2011). Transportation Mode Detection using Mobile Phones and GIS Information. *19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 54–63.

Stopher, P. (2008). Search for a global positioning system device to measure person travel. *Transportation Research Part C: Emerging Technologies*, 350-369.

Tennøy, A., & Knapskog, M. (2022). Walking distances to public transport in smaller and larger Norwegian cities. *Transportation Research Part D: Transport and Environment*.

Ting Zuo, H. W. (2018). Determining transit service coverage by non-motorized accessibility to transit: Case study of applying GPS data in Cincinnati metropolitan area. *Journal of Transport Geography*, 1-11.

Xiao, G. (2015). Travel mode detection based on GPS track data and Bayesian networks. *Computers, Environment and Urban Systems*, 14-22.

Yazdizadeh, A. (2019). Ensemble convolutional neural networks for mode inference in smartphone travel survey. *IEEE Transactions on Intelligent Transportation Systems*.

Yu, J. J. (2020). Semi-supervised deep ensemble learning for travel mode identification. *Transportation Research Part C: Emerging Technologies*, 120-135.

Zheng, Y. (2011). *Geolife GPS trajectory dataset.* Microsoft Research Asia.