Assessing the Potential of Google Location History (GLH) Data for Travel Behavior Research in the Context of Developing Country*

Kaiser Hamid, Md. Sayem Noor and Annesha Enam, PhD

***Abstract*— Using passive data to investigate travel behavior is becoming increasingly prevalent, owing to its convenient data acquisition process. This study seeks to evaluate the feasibility of leveraging Google Location History (GLH) data for analyzing travel behavior within the context of a developing nation like Bangladesh, characterized by high population density, diverse land use, and heterogeneous traffic patterns, including a significant presence of non-motorized vehicles, and relatively low motorized vehicle speeds. A group of 60 individuals willing to share their GLH data stored in the Google Maps application was recruited to accomplish this. A dedicated mobile phone application named Trip Tracker was developed to facilitate the collection of ground truth data.**

Validation of the GLH data was carried out through a three-step procedure. Initially, the identification of home and work locations from GLH, based on visit frequency and duration, was cross-verified against user-provided inputs, demonstrating 100% accuracy. Subsequently, the accuracy of day-to-day travel data, including arrival and departure times and locations, was assessed against GLH information, yielding a spatial and temporal matching accuracy of 82%. Thirdly, the modes of transportation extracted from ground truth data were compared with those provided by GLH, revealing a mode prediction accuracy of 53% for GLH data. This discrepancy was attributed to the intricate nature of Dhaka's traffic system and the prevalence of non-motorized transportation modes like rickshaws. Additionally, GLH tends to aggregate multimodal trips, revealing only the high-speed mode and neglecting the mode(s) used for the last/first-mile connection.

Consequently, two predictive models were developed utilizing Random Forest (RF), a tree-based machine learning (ML) algorithm, and a long short-term memory neural network (LSTM-based NN) to refine the GLH-predicted travel mode information. The RF and LSTM models achieved mode prediction accuracies of 86% and 68%, respectively, representing a notable improvement over GLH predictions. Further enhancements in accuracy can be anticipated by increasing the sample size.

I. Introduction

The traditional method of travel data collection entails active solicitation in the form of in-person, telephone, or mail-back interviews, such as the decennial National Household Travel Survey (NHTS) of the USA [1] and the National Travel Survey of the UK [2]. However, such data collection technique has met with several condemnations, including data quality issues such as missing trips, particularly by non-motorized vehicles, and excessive respondents' burden, especially while collecting multiday activity-travel information. Moreover, relying on active solicitation for data collection could be prohibitively challenging for developing countries due to budget constraints, institutional weakness, and lack of skilled personnel.

As a result, the use of passive data such as those collected from global positioning system (GPS) [3-6], social media check-in [7], smart transit fare cards [8], mobile phone call data [9-10], mobile phone location history [11] is garnering popularity, especially during the last two decades. For example, data obtained from smart subway fare card transaction information were successfully utilized by Hasan et al. [8] to depict urban mobility patterns. Similarly, Hasan et al. [7] used social media check-in data to reveal the impact of social influence on people's choices and lifestyles. Toader et al. [6] utilized GPS data to analyze the activity preferences of individuals. Cantelmo et al. [11] extended the study by employing mobile phone location history to automatically detect activity locations and eliminate the need for direct interactions with the participants.

More recently, researchers have started to explore the plausibility of using Google location history (GLH) data provided by the Google Map smartphone application (App) to infer human mobility patterns [12]. Cools et al. [13] evaluated the feasibility of using GLH data to substitute travel diary information. They investigated the GLH data's effectiveness in detecting locations and trips across diverse urban environments. However, they discovered that Google Maps often fails to account for locations with shorter dwell times. Another study used GLH data to assess its potential to detect the joint activities of the traveler [14]. The studies above show the possibility of using GLH for travel behavior research. However, none of the studies investigated the performance of the GLH data for inferring important travel information such as trip distance, travel time, and trip mode in the context of developing countries, which is often characterized by high population density, mixed use of land, heterogeneous traffic, including a high share of non-motorized vehicles and low speed of motorized vehicles.

Therefore, the purpose of the current study was to assess the pertinence of GLH data extracted from the Google Map smartphone application (App) for conducting travel behavior research in Dhaka, Bangladesh – one of the most populated megacities of South Asia that suffers from a lack of periodic

*Research supported by CASR, BUET.

Kaiser Hamid is a Graduate student at Bangladesh University of Engineering and Technology, Dhaka-1000 (email: 1804083@ce.buet.ac.bd).

Md. Sayem Noor is a Graduate student at Bangladesh University of Engineering and Technology, Dhaka-1000 (email: 1804037@ce.buet.ac.bd).

Annesha Enam, PhD, is an Associate Professor at Bangladesh University of Engineering and Technology, Dhaka—1000 (email: annesha@ce.buet.ac.bd), and the paper's corresponding author.

travel survey for capturing the travel trends of its expanding population. To this end, the study recruited a panel of participants willing to share the GLH data stored in their Google Map application. The participants also logged their day-to-day travel information through a mobile phone application (App) – Trip Tracker developed for this study. The data collected using Trip Tracker – often referred to as the ground truth data or offline data in the paper – was used to check the performance of the GLH data for inferring pertinent activity-travel information.

The validation of the GLH data was accomplished in three steps. First, the home and work locations were inferred from GLH based on the frequency of place visits and duration of stay and were validated against the user input. Second, the day-to-day travel data, such as the arrival and departure times and locations, were checked against the GLH information. Thirdly, the ground truth travel modes were examined against those obtained from Google Location History (GLH). Preliminary analysis revealed that GLH's prediction of Dhaka's travel mode was unsatisfactory. It was found that GLH systematically misclassifies several popular local modes of Dhaka. For example, rickshaws, e-rickshaws, CNG, and tempo are often inaccurately classified as passenger vehicles or buses. Additionally, Google does not label the multimodal trips correctly. Instead, it reveals only the high-speed mode from the sequence of modes and neglects the modes used for first and last-mile connectivity.

Consequently, the study developed a predictive model for travel mode classification using Random Forest (RF), a widely used decision tree-based machine learning algorithm (ML) for travel behavior modeling [15] and a long-short-term memory neural network (LSTM-based NN), capable of handling sequential data [16]. These models were constructed based on the traveler's location information provided by GLH. To the authors' knowledge, this is the first travel mode prediction model for Dhaka based on the semantic information collected from GLH.

The rest of the paper is organized as follows: the next section describes the study methodology, section III provides the analysis of the results, and section IV concludes the paper with a summary of the study and the direction for future research.

II. METHODOLOGY

Google Location History (GLH) is a Google service that records user location data, including geocoordinates, timestamps, place names, and inferred transportation modes, when the "Location History" setting is activated in a user's smartphone Google Maps application. The study involved data collection, data analysis, validation of GLH information, and development of a mode prediction model. The following subsections provide detailed descriptions of these phases.

A. Development of Trip Tracker Application

An application (App) called Trip Tracker was developed to collect travel information such as the arrival and departure times, the origin and destination of the trips, and the travel modes of the respondents. Participants could indicate the start and end of trips through push buttons in the AppApp. A dropdown list was provided to select the transportation mode during the trip. The AppApp could capture the geocoordinates and timestamps of the starting and ending locations of the trip. Upon completing the trip, the AppApp saved the trip data in a MongoDB [17] database, associating it with the user's ID.

B. Participant Recruitment

The study started with a recruitment survey to enlist willing participants. The eligible participants were enlisted for three months. The enrolled participants provided their home, workplace, or school location during the survey. They consented to use the Trip Tracker Android app to log their travel details during the study period. The participants were requested to keep their location services active throughout the survey to collect GLH information. Throughout the survey, the study team periodically contacted the participants to ensure consistent logging of the travel information. The participants who diligently participated in the study by carrying the Trip Tracker App during all travels and handed over the GLH data were compensated for their valuable time with either a food voucher or cash payment based on their preference. The collected data were stored in a MongoDB [17] open-source database.

C. Preprocessing of GLH Data

The GLH data is obtained from the users in the form of a JSON file, in which "activitySegment" and "placeVisit" are nested in the "timelineObjects." The "place visit" segment contains information about the latitude and longitude, start and end times of a place visit, the inferred name of the visited place, and the confidence level of the inferred location. The "activitySegment" contains the start and end latitude and longitude of a trip, start and end times, the inferred mode of travel, and the latitude and longitude of some intermediate points between the trip start and end point. The "placeVisit" and "activitySegment" information are merged based on the place visit's start time and the trip's end time. The unique user ID generated during the Trip Tracker survey integrated the GLH information with the Trip Tracker Application data.

D. Feature Extraction from GLH data

The geolocation and time stamp information provided by GLH were used to calculate the trip duration, travel distance, velocity, and acceleration of the trip segments. The geocoordinates provided by GLH can be represented by the set $G = \{(\lambda_0, \phi_0), (\lambda_1, \phi_1), (\lambda_2, \phi_2), \dots (\lambda_n, \phi_n)\}$ where λ_i is i'th location's longitude and ϕ_i is the i'th location's latitude. The timestamps can be represented by $T = \{t_0, t_1, t_2, \dots t_n\}$, where t_i is the timestamp corresponding to the i'th location. The distance between two consecutive geolocations d_i was calculated using the Haversine formula [18] based on the corresponding latitude, longitude set, *G*. The trip duration *tt* was calculated using (1) from the timestamp dataset, *T* as follows

$$tt = t_n - t_o \tag{1}$$

The velocity of the trip segments was then calculated from the corresponding distance and time intervals. It can be noted that the term trip segment is used to refer to the portion of a trip between any two consecutive locations.

E. Validation of home and work location

The concept of "home" typically refers to where individuals reside at night, generally between 10 p.m. and 7 a.m. Subsequently, individuals proceed to work or school, spending a substantial period before venturing into

recreational and maintenance activities at home or outside. The study applied heuristics to identify the home and the second most frequent destination (most often work or school) from the GLH data. The paper identifies the most frequently visited location between 10 p.m. and 7 a.m. as home. Similarly, the most visited place between 7 a.m. and 5 p.m. on weekdays was identified as work/school. The identified home and work/school geolocation was checked against the user-provided input during the recruitment survey.

F. Validation of trip information

The assessment of the daily travel information relied on computing the discrepancies between start locations and end locations and start times and end times of the trips. First, the GLH and Trip Tracker data were sorted by the user ID and the trip's start time. The validation process utilized the Mean Deviation and Root Mean Square Error (RMSE) for assessment. It can be noted that 100% accuracy is obtained only when the exact geolocation (i.e., 0 m error in space) is reported by the GLH and Trip Tracker at the same time (i.e., 0 min discrepancy in time). Therefore, allowing a temporal and spatial threshold can improve the accuracy of the match. The study used a temporal threshold of 0 to 5 minutes and a spatial threshold of 0 m to 500 m.

G. Mode Split

During the quality check of the GLH data, it was noted that the GLH does not accurately report modes for multimodal journeys – instead, it reports a single mode for the whole trip of such journeys. However, multimodal trips accounted for 26% of the surveyed trips. Therefore, a trip-splitting algorithm was implemented based on the segment velocities calculated from GLH-provided location information to identify the trip legs corresponding to different modes. Typically, the velocity changes when the respondents switch their travel modes. Therefore, velocity clusters were identified along the trip segments to locate the point of modal shift. Later, the mode prediction algorithms (described next) were applied to predict the travel mode corresponding to each trip segment. This approach enabled the identification of multimodal journeys and the labeling of multiple modes of that journey.

The pseudocode for the mode split algorithm is shown in Fig. 1. It can be noted that the mode split algorithm was applied to all the trips reported via GLH. Hence, it was challenging to set the appropriate velocity threshold for clustering. The algorithm had the potential to break the singly modal journeys into smaller segments as well. The mode-splitting task involved clustering similar velocity profiles; however, the exact number of mode changes within a trip was uncertain. Two types of thresholds were employed for clustering to address this uncertainty. Initially, velocities were categorized into two groups: those exceeding 16 km/hr (considered high) and those below 16 km/hr (considered low). Subsequently, a low threshold of 3 km/hr was applied to velocities below 16 km/hr, while a high threshold of 16 km/hr was applied to velocities bel to velocities above 16 km/hr. This segmentation process partitioned the velocity list into distinct lists, each representing a transportation mode.

H. Mode Prediction Models

Two mode prediction models were developed utilizing ML based on Random Forest (RF) and the Long Short-Term Memory (LSTM) approach within an artificial neural network.

```
def identify_modes(speeds, low=3, high=16):
    modes = []
    current = [speeds[0]]
    for i in range(1, len(speeds)):
        threshold_forward = low if speeds[i] <= 16
else high
        threshold_backward = low if speeds[i - 1] <=
16 else high
        if abs(speeds[i] - speeds[i-1]) <=
threshold_forward or abs(speeds[i] - speeds[i-1]) <=
threshold_backward:
            current.append(speeds[i])
        else:
            modes.append(current)
            current = [speeds[i]]
    modes.append(current)
    return modes
```

Figure 1. Pseudocode for mode prediction model

RF is a widely used decision tree-based ensemble ML algorithm [19], capable of effectively handling high dimensional data [20], missing values and outliers [19], and class imbalance [21]. The selection of LSTM is motivated by its ability to capture intricate patterns within sequential data [16] effectively. The LSTM architecture encompasses three pivotal components: the input gate, forget gate, and output gate, in addition to the memory cell blocks relating to long and short-term memory.

Inputs for Prediction Model:

Our research employed the following seven features to construct the mode prediction model.

Segment velocity: This feature encompasses sequential data, representing the velocity at various time points during a trip.

Start time of the Trip: This feature denotes the trip's initiation time and is obtained from the GLH dataset.

Day of the week: This feature signifies the day the trip commenced.

Trip mode predicted by GLH: This feature represents the Google-predicted trip mode.

Segment time duration: This feature contains the duration of the subsequent time stamps.

Average velocity: This feature represents the average segment velocities.

Trip distance: This feature indicates the distance obtained from the final and initial geolocation of the trip.

Hyperparameters used in the RF mode prediction model:

One hundred decision trees were developed to build the random forest. The selected minimum leaf size was 1, necessitating a minimum of 2 samples per leaf. The minimum samples required for a split was set at 2. Lower values of this parameter enable the tree to capture more intricate patterns in the data, which is beneficial but can also lead to overfitting, particularly in datasets with a high level of noise. Conversely, higher values promote the development of simpler tree structures but can lead to underfitting if the values are too high [22]. Therefore, the hyperparameters were selected methodically based on a grid search cross-validation process.

Hyperparameters used in the LSTM mode prediction model:

The input layer of the proposed neural network model was divided into four branches to process distinct data types. The first branch used an LSTM to handle sequential data such as segment velocity and segment time duration. In contrast, the other five branches, dealing with time, day, Google-predicted trip mode, trip distance, and average velocity, included a dense layer with dropout. The *Relu* activation function was applied at the input layer. Outputs from all branches were concatenated and directed to a hidden layer that employed the *Relu* activation function with dropout to prevent overfitting. The output layer used the *softmax* activation function for the multi-class classification.

Performance analysis of the prediction models:

Confusion matrices were produced to compare the performance of the mode prediction models. The confusion matrix is created with the actual modes on the y-axis and the predicted modes on the x-axis. Therefore, diagonal elements of the matrix represent accurate predictions, and the off-diagonal elements represent incorrect predictions.

The accuracy of a mode, i, was calculated using (2) as follows:

$$A_i = \frac{CP_i}{N_i} \quad (2)$$

Where A_i represents the accuracy for mode, i, CP represents the correct prediction, and N represents the actual number of modes in the ground truth data.

The overall accuracy across all modes was calculated as the weighted average of the mode-specific accuracy using (3) below:

$$OA = \frac{\sum_i CP_i N_i}{\sum_i N_i} \quad (3)$$

III. Result and Discussion

Three thousand six hundred twenty trips were reported in the GLH data by the 60 respondents. A rigorous data quality check (QA/QC) was conducted before the data validation and mode prediction.

A. Validation of home and work location

The validation of work and home locations was conducted by comparing offline data with GLH data. The heuristic algorithm (mentioned in the methodology section), based on the time of day and frequency of visits, accurately predicted the home and work locations for all 60 users.

B. Validation of trip information

The origin and destination location and the arrival and departure times of all GLH trips were compared against the data collected via the Trip Tracker application. The data comparison resulted in an 82% accuracy while considering a space threshold of 500 m and a time threshold of 5 minutes. In other words, for 82% of the trips, the origin and destination locations reported in the GLH data and by the Trip Tracker were within 500m of each other, and the arrival and departure times of these trips were within 5 minutes. The mean deviation and the RMSE value of the origin, destination, departure, and arrival times for these 2960 trips (82% of 3620) trips are reported in Table I.

TABLE I. THE MEAN DEVIATION AND RMSE BETWEEN GLH DATA AND OFFLINE DATA

Parameters	Mean deviation	RMSE
Origin	0.0449 km	0.105
Destination	0.038 km	0.090
Departure time	0.71 minute	1.29
Arrival time	0.81 minute	1.37

C. Validation of travel mode

The 2960 trips validated for spatial and temporal accuracy were used further to validate travel mode. While comparing GLH with the offline data, it was found that GLH cannot capture all the trips undertaken by the respondents. Fig. 2 shows the distribution of transport modes not identified by GLH. The values in parentheses represent the average length of the missed trips. As shown in Fig. 2, walking accounts for more than 50% of the transport modes missed by GLH. Further analysis revealed that the average length of the trips missed by GLH is less than 500m. The frequent omission of walking trips by GLH can be attributed to their low velocity and smaller radius of movement – a finding that aligns with observations of Cools et al. [13].

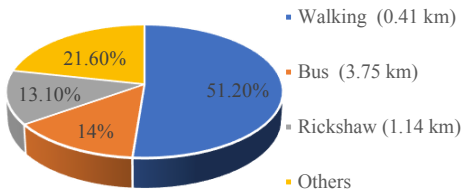

Figure 2. Most frequent missing trip mode and the associated distance

Next, a confusion matrix was created to compare the GLH-predicted modes against the modes reported by the respondents in the offline data. Fig. 3 presents the confusion matrix. From the confusion matrix, it is apparent that approximately 70% of the reported bike and bus trips are classified as motorcycle and bus, respectively, by GLH. Moreover, GLH does not use separate categories for CNG, Tempo, and E-Rickshaws; instead, it categorizes these vehicles as passenger vehicles, buses, walking, or subways. Nonetheless, the GLH labels the cycling and the walking trips satisfactorily. The overall mode prediction accuracy for GLH in Dhaka was only 53%. Additionally, the multimodal journeys are represented by a single mode in the GLH data.

D. Mode prediction model results

80% of the 2960 validated trips were used to train the RF and LSTM-based neural network models, and 20% of those trips were used to test the predictive accuracy of the developed models. It may be noted that CNG trips were excluded due to their insufficient number of observations. Table II compares the predictive accuracies of the RF and LSTM-based neural network models. It is evident from the table that the RF model attains a higher overall accuracy of 86%, while the LSTM model achieves an accuracy of 68%. However, both models

surpass the 53% accuracy rate obtained using GLH data. The relatively lower accuracy of the LSTM model, as compared to the RF model, could be attributed to the insufficient quantity of data available for training the neural network and the presence of class imbalances.

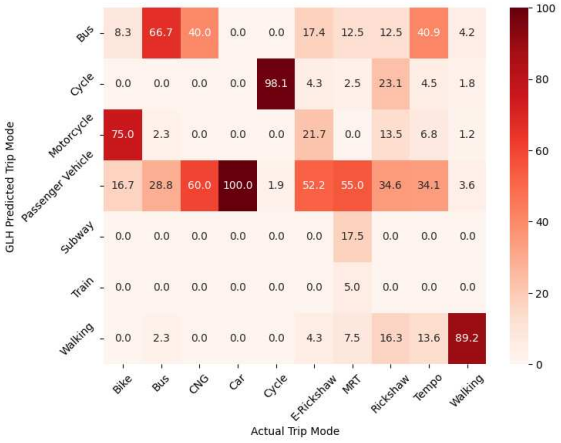

Figure 3. Confusion matrix for GLH predicted mode and actual mode

TABLE II.Comparison of RF and LSTM mode prediction model

Modes	**Mode Specific Accuracy for RF and LSTM model**			
	Accuracy for RF model	***Accuracy for LSTM model***	***No. of the test sample***	***No. of training sample***
Bike	1.00	0.85	79	316
Bus	0.81	0.67	78	312
Car	0.78	0.57	65	260
Cycle	0.90	0.60	69	276
E-Rickshaw	0.82	0.36	39	156
MRT	0.95	0.85	39	156
Rickshaw	0.60	0.58	80	320
Tempo	1.00	0.51	37	148
Walking	0.93	0.86	106	424
Overall Accuracy	0.86	0.68	592	2368

Fig. 4, which displays the RF model's confusion matrix, indicates accurate bike and tempo predictions. However, it shows a reduced accuracy of 60% for rickshaws, as delineated in Table II. In most instances, the rickshaw is incorrectly predicted as a cycle. Conversely, the LSTM model exhibits its highest accuracy at 86% for walking, while its accuracy for predicting e-rickshaws is the lowest at 36%, as shown in Table II. According to the confusion matrix in Fig. 5, e-rickshaws are predominantly misclassified as rickshaws in the LSTM model.

Therefore, the proposed RF model can predict the transportation modes for Dhaka sufficiently accurately using the location information provided by GLH. This is the first attempt to identify Dhaka's heterogeneous travel modes using the limited information provided by GLH to the best of the author's knowledge.

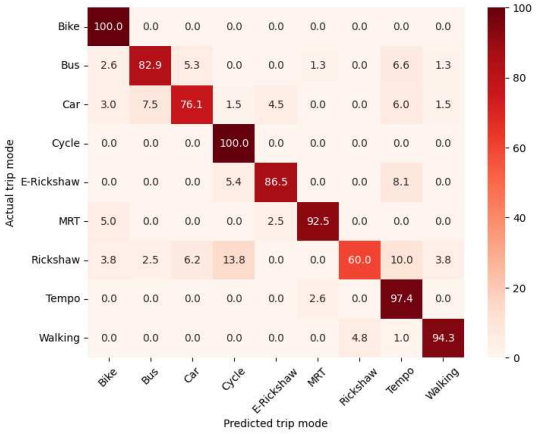

Figure 4. Confusion matrix from the RF model

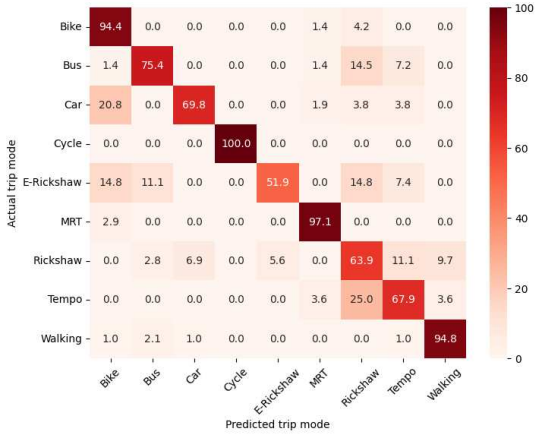

Figure 5. Confusion matrix from the LSTM model

IV. CONCLUSION

This study evaluates the potential of Google Location History (GLH) data for travel behavior research in Dhaka, a city characterized by a heterogeneous traffic stream and unbearable traffic congestion. The key findings of the study can be outlined as follows.

The study successfully identified the respondents' home and work/school location using the GLH data based on the frequency of visits (heuristic) algorithm. The spatial (origin and destination) and temporal information (arrival and departure time) of 82% of the GLH trips matched those collected using the Trip Tracker application.

However, the accuracy of the GLH inferred travel modes in the context of Dhaka was not found to be satisfactory. For example, GLH was found to mislabel motorcycles or motorbikes as bikes. Similarly, autorickshaws such as CNG, tempo, and battery-driven electric rickshaws, quite prevalent in Dhaka [23], are mislabeled as passenger cars, public transit, or walking. Additionally, GLH systematically mislabels the multimodal journeys, reporting only one mode and neglecting the others.

The paper proposed a heuristic algorithm to split the multimodal trips into constituent travel modes. This additional

step was necessary since GLH was found to report a single mode for multimodal trips. Next, the paper proposed ML-based RF and LSTM-based neural network models trained on GLH-provided location information. The RF and the LSTM-based NN model produced a mode prediction accuracy of 86% and 68%, respectively, significantly higher than the accuracy provided by GLH for Dhaka.

The higher accuracy obtained in the current paper is a noteworthy achievement since this is one of the first attempts to classify travel modes from the limited information provided by GLH in the context of heterogeneous traffic to the best of the author's knowledge. However, the prediction accuracy could be further improved with an increased sample size. Subsequent research should examine the spatial transferability of the findings obtained in the current endeavor.

ACKNOWLEDGMENT

The corresponding author would like to acknowledge a grant from the Committee for Advanced Studies and Research (CASR), BUET.

REFERENCES

[1] Oak Ridge National Laboratory, "National Household Travel Survey (NHTS)", [Online]. Available: https://nhts.ornl.gov. [Accessed: Apr. 30, 2024].

[2] Gov. UK, "National Travel Survey Statistics", [Online]. Available: https://www.gov.uk/government/collections/national-travel-survey-statistics. [Accessed: Apr. 30, 2024]

[3] N. Shoval, M. Isaacson, and P. Chhetri, "GPS, smartphones, and the future of tourism research," *The Wiley Blackwell Companion to Tourism*, 2024, pp. 145-159

[4] P. Sivalingam, D. Asirvatham, M. Marjani, J. A. I. S. Masood, N. K. Chakravarthy, G. Veerisetty, and M. T. Lestari, "A review of travel behavioral pattern using GPS dataset: A systematic literature review," *Measurement: Sensors*, vol. 101031, 2024

[5] L. Shen and P. R. Stopher, "Review of GPS travel survey and GPS Data-Processing Methods," *Transport Reviews*, vol. 34, no. 3, pp. 316–334, 2014.

[6] B. Toader, G. Cantelmo, M. Popescu, and F. Viti, "Using passive data collection methods to learn complex mobility patterns: an exploratory analysis," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE*, 2018, pp. 993–998.

[7] S. Hasan, S. V. Ukkusuri, and X. Zhan, "Understanding social influence in activity location choice and lifestyle patterns using geolocation data," *Frontiers in ICT*, vol. 3, 2016.

[8] S. Hasan, C. M. Schneider, S. V. Ukkusuri, and M. C. González, "Spatiotemporal Patterns of Urban Human Mobility," *Journal of Statistical Physics*, vol. 151, pp. 304–318, 2013.

[9] A. Gregg, J. Blasco-Puyuelo, R. Jordá-Muñoz, I. M. Martínez, J. Burrieza-Galán, and O. C. Ros, "Airport accessibility surveys and mobile phone records data fusion for the analysis of air travel behavior," *Transportation Research Procedia*, vol. 76, pp. 269-282, 2024.

[10] Z. Patterson and K. Fitzsimmons, "DataMobile: Smartphone Travel Survey Experiment," *Transportation Research Record*, vol. 2594, no. 1, pp. 35–43, 2016.

[11] G. Cantelmo, P. Vitello, B. Toader, and F. Viti, "Inferring urban mobility and habits from user location history," *Transportation Research Procedia*, vol. 47, pp. 283–290, 2020.

[12] X. Yu et al., "On the accuracy and potential of Google Maps location history data to characterize individual mobility for air pollution health studies," *Environmental Pollution*, vol. 252, pp. 924–930, 2019.

[13] D. Cools, S. C. McCallum, D. Rainham, N. Taylor, and Z. Patterson, "Understanding Google location history as a tool for travel diary data acquisition," *Transportation Research Record*, vol. 2675, pp. 238–251, 2021.

[14] G. Parady, K. Suzuki, Y. Oyama, and M. Chikaraishi, "Activity detection with Google Maps location history data: Factors affecting joint activity detection probability and its potential application on real social networks," *Travel Behaviour and Society*, vol. 30, pp. 344–357, Jan. 2023.

[15] L. Cheng, X. Chen, J. De Vos, X. Lai, and F. Witlox, "Applying a random forest method approach to model travel mode choice behavior," *Travel Behaviour and Society*, vol. 14, pp. 1-10, 2019.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] MongoDB, Inc., "MongoDB", [Online]. Available: https://www.mongodb.com [Accessed: Apr. 30, 2024]

[18] K. Gade, "A non-singular horizontal position representation," *The Journal of Navigation*, vol. 63, no. 3, pp. 395–417, 2010.

[19] L. Breiman "Random forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[20] G. Biau, "Analysis of a random forests model," *J. Mach. Learn. Res.*, vol. 13, pp. 1063-1095, 2012. [Online]. Available: http://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf

[21] A. S. More and D. P. Rana, "Review of random forest classification techniques to resolve data imbalance," in *Proc. 2017 1st Int. Conf. Intelligent Systems and Information Management (ICISIM)*, 2017, pp. 72-78

[22] P. Probst, M. N. Wright, and A.-L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 3, Art. no. e1301, 2019

[23] H. Mohiuddin, M. M. R. Bhuiya, M. M. U. Hasan, and H.-T. Jamme, "How individual perceptions of transportation systems influence mode choice for mobility-challenged people: a case study in Dhaka using an integrated choice and latent variable model," *Transport Policy*, vol. 147, pp. 259-270, 2024.