Exploring the Impact of Varying GPS Trajectory Frequencies and Class Balancing on Travel Mode Detection

Reem ALOLABI[1], Fahima Lailul ULA[2], Varun VARGHESE[3] and Makoto CHIKARAISHI[4]

[1]Graduate School of Advanced Science and Engineering, Hiroshima University
(1-5-1 Kagamiyama, Higashi-Hiroshima City, Hiroshima, 739-8529, Japan)
E-mail: reem98olabi@gmail.com
[2]Graduate School of Advanced Science and Engineering,Hiroshima University
(1-5-1 Kagamiyama, Higashi-Hiroshima City, Hiroshima, 739-8529, Japan)
E-mail: fahimalula@gmail.com (Corresponding Author)
[3]Member of JSCE, Assistant Professor, Graduate School of Advanced Science and Engineering,Hiroshima University
(1-5-1 Kagamiyama, Higashi-Hiroshima City, Hiroshima, 739-8529, Japan)
E-mail: varunv@hiroshima-u.ac.jp
[4]Member of JSCE, Professor, Graduate School of Advanced Science and Engineering,Hiroshima University
(1-5-1 Kagamiyama, Higashi-Hiroshima City, Hiroshima, 739-8529, Japan)
E-mail: chikaraishim@hiroshima-u.ac.jp(CorrespondingAuthor)

In recent years, machine learning models using big data such as GPS have gained popularity in travel mode detection. However, GPS data collection consumes high energy in smartphones, impacting battery life. This could be handled through lowered sampling of GPS points but its impact on accuracy and interpretability are not well understood. Another issue in travel mode detection arises from the imbalance between different modes, resulting in lower accuracy for minority modes. Using GPS data collected from Hiroshima, Japan, this study quantifies the relationship between accuracy and GPS sampling interval, while quantifying the interpretability of different machine learning algorithms. Our empirical results show that the accuracy of imbalanced minority modes are lower consistently through all scenarios but improves when the data is balanced. In terms of interpretability, GPS related variables such as speed become less important when fewer GPS points are sampled, while GIS and GTFS related variables such as buffer ratio near bus stops become more important. Overall, our results show that, for effective implementation in practical settings, it becomes crucial to adapt feature engineering and model types based on the available data type.

Key Words: *Mode Detection, GPS Trajectory Data, Class Imbalance, GPS sampling interval*

1. Introduction

Travel mode detection plays a crucial role in shaping transportation planning policies. Accurately predicting travel modes is vital for developing efficient transport-related policies and services, including traffic management, sustainable transportation systems, demand management, and multi-modal integration. In recent years, machine learning models have gained popularity for travel mode detection due to their higher accuracy compared to theoretical discrete choice models (Varghese, Chikaraishi, Urata, et al., 2020). However, interpretability of these models remains a challenge. Additionally, an important issue in travel mode detection arises from the imbalance between different modes. This imbalance is primarily caused by uneven levels of transport service provisions, leading individuals to prefer a particular mode (known as the majority mode) over other modes (known as minority modes) (Chen & Cheng, 2023). The class imbalance problem results in lower accuracy levels of machine learning models, especially for minority modes (Abd Elrahman & Abraham, 2013).

The popularity of smartphones has led to an increased use of Global Positioning System (GPS) data in transportation research. GPS trajectories provide time-series data with multiple points, each containing latitude, longitude, and timestamp information. Unlike traditional travel diary surveys, GPS data reduces the burden on respondents as it does not require them to actively report details of each trip. Nonetheless, GPS trajectory data has two significant limita-

tions (M. Lin & Hsu, 2014). First, systematic errors exist in GPS data, especially in underground areas, indoors, and near tall buildings, resulting in poor data quality. Second, GPS data collection consumes high energy in smartphones, impacting battery life. Although the trade-off between accuracy and energy consumption has been acknowledged (K. Lin, Kansal, Lymberopoulos, & Zhao, 2010; Nawarathne et al., 2014), limited attention has been given to this limitation in the transportation literature. Efforts have been made to address energy efficiency in GPS data collection, such as phased sampling to turn off GPS for extended periods or reducing the number of sampled GPS points in smartphone applications(Cottrill et al., 2013; Marra, Becker, Axhausen, & Corman, 2019). However, the actual impact of varying sampling frequencies on prediction accuracy and model interpretability in transportation remains rarely quantified. Additionally, for GPS data, the class imbalance issue is linked to the data frequency, and varying sampling frequencies can significantly impact the accuracy and interpretability of machine learning models. However, the interaction between varying GPS sampling frequencies, class imbalance, and model performance has not been extensively explored in the literature.

(1) Relevant Literature

The high battery consumption of GPS-based applications have been long identified and various travel survey applications and transportation researchers have tried to address this issue. Studies discuss different sampling strategies such as periodic sampling and phased sampling employed in travel survey applications such as Future Mobility Survey (FMS)(Cottrill et al., 2013; Nawarathne et al., 2014). They observed that detection accuracy decreased with longer sleeping durations in phased sampling and longer sampling periods in periodic sampling. Furthermore, they noted a generally low detection accuracy for Mass Rapid Transit (MRT) trips due to poor GPS signal reception inside the MRT system. Other studies, such as Xiao et al. (Xiao, Juan, & Zhang, 2016) and Maruyama et al. (Maruyama, Sato, Nohara, & Imura, 2015), provided survey participants with a power bank and additional batteries, respectively, to collect GPS trajectory information. Researchers realize that there exists a trade-off, however, the thresholds employed to save energy are decided without any clear reasoning. The impact of GPS location data frequency on the accuracy of prediction in transportation planning variables is not clearly understood. Additionally, there is not much evidence available on how class-imbalance between travel modes and varying GPS data frequency influence mode detection accuracy.

Furthermore, there has been an over reliance on accuracy as the sole indicator to assess the effectiveness of detection algorithms, with relatively fewer studies exploring the interpretability of these algorithms. Recently, Shapley additive explanations (SHAP) values based variable importance (Lundberg & Lee, 2017) have been evaluated in travel mode prediction studies(Tamim Kashifi, Jamal, Samim Kashefi, Almoshaogeh, & Masiur Rahman, 2022). SHAP values provide insights into each variable's partial contribution, and these values can be used to calculate relative variable importance. Variable importance measures are often recommended as a means to select the best features for model development. However, little is known about how data collection frequency and class imbalance impact these variable importance measures.

(2) Research Objectives

This study aims to achieve the following three research objectives. *Firstly*, to analyze the effect of varying GPS trajectory frequencies on the prediction accuracy of travel mode detection. By varying the frequency of GPS trajectories i.e., increasing the average time between point sampling, this study aims to quantify the impact of reducing the sampling of GPS points on prediction accuracy. *Secondly*, to analyze the impact of varying GPS trajectory frequencies on the contribution of different variables using interpretable machine learning models. Estimating SHAP values will counter the black-box functionality of machine learning models and analyze how the relative importance of variables gets affected by GPS data frequency. *Thirdly*, to address the issue of class imbalance when evaluating the effect of varying GPS sampling frequency on model accuracy and interpretability. The empirical analysis utilizes GPS trajectories collected from smartphone-based travel surveys conducted in three waves in Hiroshima prefecture, Japan. During the survey, respondents were prompted to provide information on trip start, trip end, travel mode, and travel purpose, while the application recorded GPS trajectories in the background. The availability of the labeled data provided by the respondents allows for testing supervised learning algorithms for travel mode detection. However, the dataset is skewed towards car and walk modes, while bicycle, bus, train, and motorcycle have relatively fewer sample points. By adjusting for this imbalance while varying GPS frequency, the study will assess the impact on model accuracy and variable importance.

The remaining sections of this article are organized as follows. In the next section, we outline the data processing technique implemented in this study to mitigate systematic errors present in the GPS data. Following that, we introduce the machine learning algorithms employed for prediction and explains the meth-

ods used to calculate variable importance. The results of the study are discussed in the subsequent section, while in the final section, the study concludes the article with a discussion on the key contributions of this work and potential avenues for future research.

2. Data processing

This section provides an overview of the data used in our study and outline the techniques used for data processing. We have two main types of data: the main trajectory data and GIS data. Additionally, we incorporate socio-economic and weather data to enrich our analysis. Figure **??** below illustrates the overall workflow encompassing data cleaning, preprocessing, and integration stages.

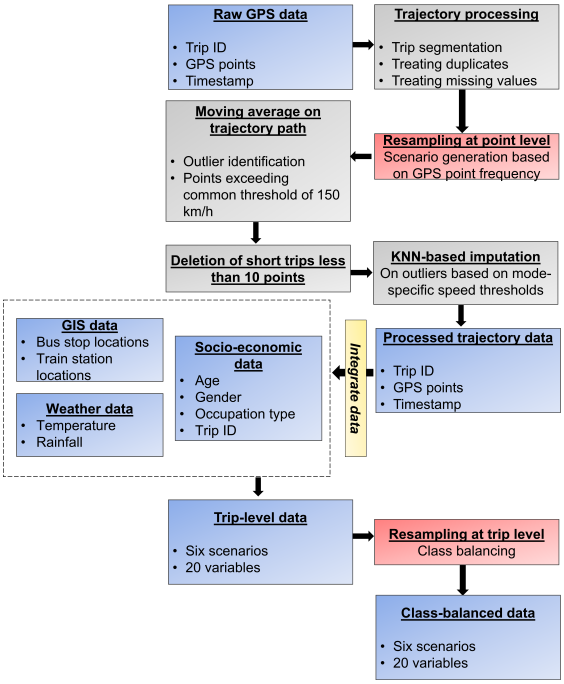


Fig. 1: **Collection of GPS Trajectory Data**

(1) Data Collection: Trajectory Data Preparation

The trajectory data was collected from travel surveys conducted through a smartphone-based travel application, covering three distinct periods: the first wave of the survey was conducted from December 2018 to January 2019 and included 5,741 trips; the second wave took place from January to February 2020 and comprised a total of 2,726 trips; and the third wave was conducted from October to November 2020, with a total of 4,374 trips. Simultaneously, socio-economic data for each user were collected as part of the three waves. Figure 2 illustrates how the trajectory data was collected. As a part of the survey, the respondents were prompted to mark the start and end of each trip, thus providing the label information on the travel mode they selected. Meanwhile, the smartphone application recorded their GPS trajectory.

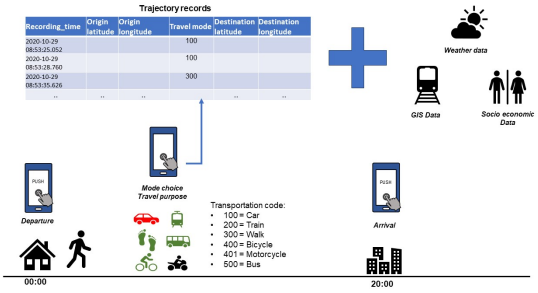


Fig. 2: **Collection of GPS Trajectory Data**

The trajectory data was collected from travel surveys conducted through a smartphone-based travel application, covering three distinct periods: the first wave of the survey was conducted from December 2018 to January 2019 and included 5,741 trips; the second wave took place from January to February 2020 and comprised a total of 2,726 trips; and the third wave was conducted from October to November 2020, with a total of 4,374 trips. Simultaneously, socio-economic data for each user were collected as part of the three waves. Figure 2 below illustrates how the trajectory data was collected. As a part of the survey, the respondents were prompted to mark the start and end of each trip, thus providing the label information on the travel mode they selected. Meanwhile, the smartphone application recorded their GPS trajectory.

The study area for this research is Hiroshima, one of the major prefectures in the Chugoku region of Japan. According to the last conducted census survey in 2020, Hiroshima prefecture had a population of approximately 2,794,862 residents (Statistics Bureau & Communications, 2021). The geographical area of Hiroshima is 8,479 km² (3,274 sq mi). Figure 3 shows a visual representation of trip nodes (i.e., origin points) in our trajectory data on the map of Hiroshima. Most of the trips were conducted primarily by two transportation modes: cars and walking, followed by a smaller number of trips via public transportation, such as buses and trains, and a few trips made by bicycles and motorcycles. The majority of the recorded trips took place in the center of Hiroshima city.

The GPS data collected in this survey had to be preprocessed to be utilized for mode detection. The primary challenge lies in the occurrence of lags when capturing departure and arrival times, caused by users forgetting to stop recording upon completing their trips. As a consequence, travel times may appear prolonged, even when the location data shows that the

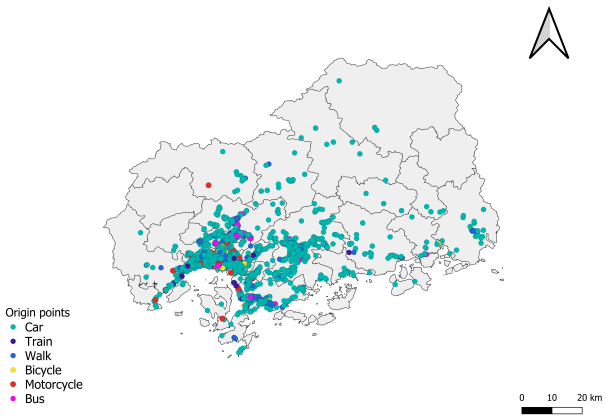


Fig. 3: **Study Area and Mode-wise Distribution of Trip Origins**

user did not cover a significant distance. Additionally, we encountered other substantial errors in capturing geolocations of departure or arrival locations due to an unstable network connection inside buildings, as previously noted in GPS data studies (Poucin, Farooq, & Patterson, 2018; Teranishi, Kimata, Kawai, & Harai, 2019). These connectivity issues led to inaccuracies in the latitude and longitude data for the traveled locations.

To address the issues caused by lags in data and improve data quality, a series of preprocessing steps were undertaken (see Figure 1). Firstly, trip segmentation was performed, where a trip was segmented into two trips whenever the user remained stationary, and a gap of over 180 seconds was detected. Subsequently, time duplicates caused by errors in the network connection were meticulously removed from the data. In addition, to further refine the geolocation information, data points with zero latitude and longitude values were removed to minimize noise in the trajectory data. Despite this process, the outliers in the trajectory data persisted. Therefore, a more detailed data processing phase was initiated, aiming to achieve a smoother data suitable this study.

The initial step before undertaking the detailed data preprocessing involved calculating the speed values from the latitude and longitude coordinates for each trajectory data point. Subsequently, we implemented a common speed threshold of 150 km/h for all GPS points. This step involved identifying and removing any trajectory data record that exceeded the specified threshold, effectively reducing noise and irrelevant data points from the dataset. Then, we filtered out the extremely short trips that consisted of less than 10 GPS points. However, outliers caused by inaccurate latitude and longitude readings still persisted. To address this challenge, the study drew inspiration from the work by Li et al. (Li, Miyashita, Kato, & Miyazaki, 2000), adopting a moving average approach with a window size of 5. This method captured underlying pattern in trajectory data by calculating the mean speed values from five consecutive data points. The resulting average value is then assigned to the central data point, within the window. As a result, the trajectory data was effectively smoothed, reducing the random fluctuations in inaccurate GPS points and improving the clarity and reliability of the location information.

As the final step, data imputation technique were applied to scrutinize any remaining speed outliers. This research utilizes specific imputation technique that is K-Nearest Neighbors (KNN) (Wolf, Loechl, Myers, & Arce, 2001; Axhausen, Schönfelder, Wolf, Oliveira, & Samaga, 2003). It involves finding the 'k' nearest data points to the outlier, where 'k' is set to 5. The outlier values are then replaced with the average or weighted average of these nearest neighbor values. For the identification of these outliers, specific speed threshold values were utilized for six transportation modes: 120 km/h for car and train, 100 km/h for bus, 97 km/h for motorcycle, 50 km/h for bicycle (including electric bicycle), 15 km/h for walk and acceleration threshold 10 m/s2 was employed as well. The idea behind selecting these thresholds was not to be accurate but be close enough to remove extreme outliers.

(2) Feature Engineering

After cleaning and processing the trajectory data, we utilized GIS data solely to extract essential features related to bus and train station stops within the Hiroshima region. The GIS data encompasses information about the geographical features and locations across the earth's surface. This enabled us to identify and incorporate valuable information regarding the origins and destinations of individual trips. Then, we integrated it with weather and the socio-economic dataset. The socio-economic data contains information about the social and economic characteristics of users, such as age, gender, and occupation type of each certain user. This data can be used to understand how travel patterns and behavior are influenced by social and economic factors. Additionally, we incorporated weather data to gain insights into the prevailing weather conditions during different trips. Understanding the impact of weather on travel patterns is vital to comprehending users' mode choices and preferences. Through the utilization of these datasets, various features have been extracted, as shown in Table 1 below.

Using the trajectory data, a total of nine features or variables were created. Initially, the trip duration (V8) feature was computed by capturing the recorded departure and arrival times, providing a precise rep-

Table 1: **Description of Data and Variables**

Code	Variable Name	Mean (Std Dev)
Trajectory data		
V1	Max speed (m/s)	17.99 (9.97)
V2	Min speed (m/s)	0.00 (0.00)
V3	Mean speed (m/s)	6.73 (4.29)
V4	Std dev speed (m/s)	3.65 (2.01)
V5	Distance (m)	6984.30 (11890.81)
V6	Mean acceleration (m/s^2)	0.095 (0.26)
V7	Std dev acceleration (m/s^2)	1.12 (0.61)
V8	Trip duration (min)	1270.31 (1273.72)
V9	Trip start time (hours)	13.28 (4.68)
GIS data		
V10	Buffer ratio for train stations	0.07 (0.15)
V11	Buffer ratio for bus stops	0.14 (0.14)
V12	Origin bus stop density	7.99 (5.70)
V13	Destination bus stop density	8.09 (5.83)
V14	Origin train station density	1.59 (2.84)
V15	Destination train station density	1.66 (2.92)
Weather data		
V16	Temperature (°C)	7.35 (5.38)
V17	Rain (mm)	0.07 (0.34)
Socio-economic data		
V18	Occupation (categories)	-
V19	Age (categories)	-
V20	Gender (female dummy)	-

resentation of the total travel time in seconds for each individual trip. Following the approach by Deniau et al. (Deniau, 2015), the distance in meters was subsequently computed using the "haversine" formula, a method available in the Python library, to determine the geographical distance between each consecutive GPS point within a trip. Subsequently, the speed in m/s for each GPS point within a trip was calculated by utilizing both the distance and the corresponding recording time at each point. Then, the acceleration was computed based on the changes in speed across consecutive GPS points. Using these measures at the point level, several variables were calculated at the trip level. They include, the mean (V3), maximum (V1), and minimum speeds (V2) (in m/s) for each trip, the mean acceleration (V6) (in m/s^2), the standard deviations in speed (V4) and acceleration (V7) in a trip. Additionally, a feature representing the distance (V5) (in m) was also included, that in principle is different from the travel time duration as the latter handles congestion and the differences in speed across modes, whereas the former might also have an important contribution in mode detection. Finally, the "trip start" time (V9) has been utilized, which shows the start hour when the trip had started. Table 1 shows the mean and standard deviation of these variables for the processes dataset in the default scenario i.e., the scenario without any resampling.

General Transit Feed Specification (GTFS) data was incorporated to investigate the influence of public transit on user mobility. As discussed in previous research on detecting user mobility accurately (Wessel & Farber, 2019), activities related to public transit play a crucial role in predicting travel behavior. To extract the GIS-related features, we followed a step-by-step process involving the conversion of trajectory data and GTFS data for stations and bus stops into a geodataframe using Python's geospatial data manipulation library. The geodataframe is a specialized data structure that allows for easy handling and analysis of geographical data.During this conversion, the latitude and longitude coordinates from the trajectory data were transformed into geometrical points. These points were then projected onto the map using the EPSG (European Petroleum Survey Group) registry code 32653. EPSG 32653 corresponds to a specific coordinate reference system that ensures precise location mapping within the Japan region. Subsequently, based on the processed geodataframe, we derived several GIS-related variables, which are listed in Table 1. Firstly, we calculated the buffer ratios for both bus stops (V11) and train stations (V10), considering a 50-meter radius around each transit station. These buffer ratios represent the proportion of GPS points recorded within the designated buffer zones to the total number of GPS points during a trip. Then, we quantified the density of bus stops and train stations at both trip origins and destinations (V12-V15, respectively). This density represents the number of transit points within a radius of 1 km^2 from the starting (origins) and ending (destinations) locations of each trip. By examining these density values, we can gain valuable insights into the availability and accessibility of public transportation options in different areas. Overall, these GIS-related features provide us with a comprehensive spatial perspective on user travel behavior and public transportation usage.

As for the weather data, two primary features, temperature and rain, were considered. The weather data includes the precipitation rate, which is indicated by the "rain" feature in inches (V17), and the temperature, indicated by the "temperature" feature (V16) in Celsius (°C) units.

Additionally, socio-economic variables such as

”age,” ”gender,” and ”occupation” were taken into account for detection. All of these variables were categorical. Age was divided into six categories: 1 = less than 20 years old, 2 = 20-29 years old, 3 = 30-39 years old, 4 = 40-49 years old, 5 = 50-59 years old, and 6 = 60 years or above. Occupation had nine categories: 1 = company employee, 2 = civil servant/organization employee, 3 = manager/executive, 4 = self-employed/free enterprise, 5 = part-time job, 6 = student, 7 = housewife (househusband), 8 = unemployed, and 9 = other. Lastly, gender had two categories: 1 = male and 2 = female. The descriptive statistics on all 20 variables (V1-V20) are presented in Table 1.

(3) Scenario Generation

To achieve the objectives of our study and thoroughly assess the performance and interpretability of machine learning models under varying levels of GPS frequencies, we employed two resampling techniques on the merged dataset: resampling at the GPS point level and resampling at the trip level.

a) Sampling Scheme at the GPS Point Level

At the GPS point level, we conducted sampling to explore the effect of varying GPS data collection frequencies. This involved creating six distinct scenarios, each with a different level of data sparsity. In the first scenario, labeled as ”No GPS points deleted,” we retained the original dataset without removing any GPS points, representing the default data with no reduction in frequency. The average time between two GPS points in this scenario was observed to 5.51 s. In the second scenario, ”Delete every 2nd point in every trip,” we selectively removed every other GPS point within each trip, effectively doubling the time interval between consecutive points (9.52 s). Similarly, in the third scenario, ”Delete 5 consecutive points in every trip,” we removed five consecutive GPS points from each trip, further reducing the frequency of data collection. This increased the average time between GPS points to 20.48 s, while substantially decreasing the total GPS points in the dataset. We continued this process, progressively removing 10 (36.01 s), 25 (74.9 s), and 50 consecutive GPS points (122.19 s) in the fourth, fifth, and sixth scenarios, respectively.

By creating these scenarios, we aimed to assess how changes in GPS data collection frequency influenced the performance and interpretability of our machine learning models. Each scenario represented a different level of data sparsity (see Table 2), with increased intervals between GPS points. This allowed us to investigate how reductions in data frequency impacted the accuracy and reliability of travel mode prediction, as well as the importance and influence of different variables within the models.

b) Sampling Scheme at the Trip Level

The resampling at the trip level was primarily conducted to address the issue of class imbalance within the dataset. Class imbalance occurs when certain transportation modes are overrepresented in the data, while others are underrepresented. This imbalance can lead to biased results and lower accuracy in the analysis, particularly for the minority modes with fewer samples. This also hinders our availability to accurately analyze the impact of varying GPS frequencies, for example if the data has two overrepresented modes such as car and walk, which are easily distinguishable based on speed-related variables, then even reducing the frequency of GPS points might not impact detection accuracy. Hence, a test must be performed on a balanced dataset.

To rectify this imbalance, we implemented a resampling scheme known as ”minimum sampling” (Susan & Kumar, 2021). This approach involved balancing the dataset by adjusting the sample size of each transportation mode to match the minimum number of trips among all modes. In other words, we randomly removed trips from the overrepresented modes until their sample sizes matched the size of the mode with the fewest trips, which in our case was the train mode. By applying minimum sampling, we ensured that each transportation mode had an equal number of representative samples, thereby addressing the class imbalance issue. This approach allowed us to create a more balanced and unbiased dataset for our analysis, improving the accuracy and fairness of the machine learning models used for travel mode prediction. Additionally, this meant that the average time in between two GPS points remained quite similar to the case when the resampling was done at the point level and thus this provides us the opportunity to test its impact on a balanced dataset (Table 2).

3. Method

This section outlines the methodology used in our research, focusing on the machine learning algorithms employed and the steps taken for hyperparameter optimization and interpretability analysis.

(1) Mode Detection Algorithms

In this study, we aim to identify transportation modes from GPS data using machine learning algorithms and explore their interpretability while varying the frequency of GPS data points. To achieve these objectives, we employed four prominent machine learning models for mode detection:

- Random Forest (RF) (Ho, 1995)
- Artificial Neural Network (NN) (Rosenblatt, 1958)

Table 2: **Sampling Strategies and Sample Sizes**

Sampling Strategy	GPS points	Av. time interval	Car	Train	Walk	Bicycle	Motorcycle	Bus
At points level								
S1: No GPS points deleted	5366901	5.51	9288	583	2555	1038	585	861
S2: Delete every 2nd point in every trip	2677576	9.52	8767	525	2124	1009	551	802
S3: Delete 5 consecutive points in every trip	1062943	20.48	7802	452	1539	869	497	689
S4: Delete 10 consecutive points in every trip	518429	36.01	6651	380	992	682	445	590
S5: Delete 25 consecutive points in every trip	184751	74.9	4452	239	352	379	342	401
S6: Delete 50 consecutive points in every trip	71003	122.19	2409	111	122	169	135	227
At trip level								
S1: No GPS points deleted	1184561	5.95	583	583	583	583	583	583
S2: Delete every 2nd point in every trip	567696	10.87	525	525	525	525	525	525
S3: Delete 5 consecutive points in every trip	216210	23.45	452	452	452	452	452	452
S4: Delete 10 consecutive points in every trip	109344	42.43	380	380	380	380	380	380
S5: Delete 25 consecutive points in every trip	37960	100.15	239	239	239	239	239	239
S6: Delete 50 consecutive points in every trip	13801	168.29	111	111	111	111	111	111

- K-Nearest Neighbor (K-NN) (Fix & Hodges, 1951)
- Support Vector Machine (SVM) (Cristianini & Shawe-Taylor, 2000)

The machine learning algorithms mentioned above have been commonly applied for transport mode detection(Cheng, Chen, De Vos, Lai, & Witlox, 2019; Byon, Ha, Cho, Kim, & Yeun, 2017; Li et al., 2021; Jahangiri & Rakha, 2015).To optimize the performance of these machine learning models, we employed ***Grid Search*** (Breiman, 2001), a hyperparameter tuning technique. Grid Search systematically explores a predefined range of hyperparameter combinations to find the optimal settings for each model. Specifically, for RF, we tuned parameters related to the *number of trees* and *maximum depth*, while for K-NN, we experimented with different *numbers of neighbors* and *distance* metrics. In the case of NN, we adjusted the learning rates and the number of hidden layers, and for SVM, we fine-tuned *regularization* and *kernel parameters*. This comprehensive hyperparameter tuning process allowed us to maximize the predictive accuracy of each model. In addition, the datasets were split into training (70%), validation (15%), and testing (15%) datasets. The grid search hyperparameter tuning was only done for the training and validation datasets and the model accuracy performance reported in the next section is based on its prediction on the testing dataset. This helped in controlling for overfitting.

(2) Interpretability

The interpretation of machine learning models is crucial in understanding the factors that contribute to model predictions and gaining insights into the underlying relationships in the data. In recent years, SHAP values based interpretability has emerged as a powerful tool for model interpretability in various domains, including transportation research. In our study we employed SHAP to assess the interpretable view of machine learning models under different scenarios and cases. The key idea behind SHAP values is to fairly distribute the *credit* or *importance* of the prediction among the input features based on their individual contributions. In essence, SHAP values assign a numerical value to each feature, indicating how much it shifts the model's prediction compared to the average prediction. Furthermore, to comprehensively evaluate model interpretability, we applied SHAP values to the RF and NN models trained on all the different sampling scenarios. Each model was trained on the datasets with varying GPS data collection frequencies and class imbalance adjustments. By understanding the contribution of each feature, transportation planners and policymakers can gain insights into the decision-making process of the models, leading to more informed and evidence-based policy decisions.

4. Results

In this section, we present the results of our study on the impact of varying GPS trajectory frequencies and class balancing on travel mode detection using machine learning models. Additionally, we delve into the interpretability of the models and discuss the implications of these findings for transportation planning and management. By analyzing the relative importance of different variables and considering class imbalance, we gain insights into the factors influencing travel mode identification accuracy and shed light on the practical applicability of the results.

(1) Algorithm comparison

The overall accuracy of the four machine learning algorithms (RF, ANN, KNN, and SVM) is presented for six different scenarios (Scenario 1 to Scenario 6) in Table 3. The accuracy in this study is denoted by the F2 score, which puts a higher focus on recall (sensitivity of true-positive), while balancing its trade-off with precision. It has been frequently observed to be the most useful metric (Hashemi et al., 2018). Along with the overall accuracy, which represents the model's ability to correctly predict the transportation mode for all classes combined, class-wise F2 scores are also presented in Table 3. We can observe that RF (F2 scores in the range of 0.87 to 0.90) and NN (0.84 to 0.86) consistently achieve the highest overall accuracy across all scenarios. In contrast, K-NN (F2 scores in the range of 0.72 to 0.79) and SVM (0.70 to 0.75) models yielded lower accuracy scores.

Furthermore, the RF and NN models demonstrated notable proficiency in correctly identifying various transportation modes compared to K-NN and SVM. For instance, in the case of the bus transportation mode, the RF and NN models achieved higher F2-scores compared to the K-NN and SVM models (see Table 3). In fact, for certain transportation modes such as trains, K-NN and SVM produced an F2 score of 0. This relatively poorer performance of these two algorithms for certain classes are consistent across scenarios. An interesting observation here is that in the imbalanced dataset, the overall accuracy is predominantly influenced by just two modes: car and walk. Due to their large sample sizes and evident differences in trajectory variables (V1-V9), they are easily distinguishable. Consequently, the model accuracy remains high even when the frequency of GPS points is reduced. Therefore, to effectively evaluate the impact of varying GPS data collection frequency on model performance, we perform class-balancing and then eval-

uate the performance of algorithms across different scenarios. However, given the poor performance of K-NN and SVM algorithms, we choose only RF and NN models for the subsequent analyses.

Class resampling or class balancing in this study was done based on minimum sampling, which adjusted the total sample size of all modes equal to the mode which had the lowest sample size. Table 4 shows the comparison between the performance of RF and NN model in both original (i.e., class imbalanced) and the sampled (i.e., class balanced) data, across all six scenarios. The results show that the overall performance of models worsen i.e., when the sample size is reduced, the models' detection ability also reduces. This is an expected finding. However, it is interesting to note that while the original imbalanced data showed no clear trends in overall accuracy of the models with changing GPS frequency, the balanced data exhibited very clear trends. The model performance decreased with a decrease in the number of sampled GPS points. In other words, as the time interval between two GPS points increased, the accuracy of the model decreased. This trend was consistent for both RF and NN models. Specifically, for the RF model, the accuracy reduced from 0.81 in scenario 1 to 0.68 in scenario 6. Meanwhile, for the NN model, the F2 score reduced from 0.64 in scenario 1 to 0.51 in scenario 6.

The trends were observed to be consistent when compared across classes i.e., the class-specific detection accuracy decrease as the frequency of GPS points decrease. Another interesting thing to note here is that class balancing reduced the difference in prediction accuracy across classes. As seen in Table 4, there was a noteworthy shift in the models' identification patterns. For instance, in scenario 1, in the original imbalanced data, car detection had an accuracy of 0.93, whereas motorcycle detection had only 0.47. However, this significant difference was not present in the balanced data, even though the overall accuracy was lower. Additionally, the detection accuracy of other road-based motorized transport modes, such as buses, also improved. Distinguishing between these road-based transport modes is particularly challenging as they exhibit similar trajectory-level features. This challenge is exacerbated when there is class-imbalanced data, as imbalanced class distributions can lead the model to favor classes with higher sample sizes, resulting in reduced accuracy for minority classes. The class resampling approach adopted in this study led to a more balanced detection of transportation modes, reducing the influence of overrepresented classes on the models' learning process. This also allowed for a clear evaluation of how variation in GPS point frequency influences model performance.

(2) Interpretability

In this section, we present the SHAP values-based interpretability analysis results for the machine learning models, namely RF and NN, in the context of transportation mode detection. The analysis is based on six different scenarios representing variations in GPS data frequency and class balancing. We aim to understand how the models make predictions, identify the key features influencing their decisions, and explore how interpretability is affected by changes in data scenarios and class resampling. Additionally, we utilized Rank-biased overlap (RBO) (Webber, Moffat, & Zobel, 2010) analysis to compare the differences in ranking across data types and scenarios. RBO is a measure commonly used to assess the similarity between two ranked lists or rankings. It takes into account the order of items in the lists and assigns higher importance to items that appear at the top of both lists, while progressively reducing the importance of items lower down the ranks.

In Table 5, the ranking of the top five features by importance for both RF and NN models in the original (imbalanced) data and the sampled (balanced) data across the six scenarios are presented. Therefore, we compare the differences in interpretability between a) model types i.e., RF vs. NN, b) scenarios i.e., the impact of GPS point frequency on interpretability, c) data types i.e., imbalanced vs. balanced data. Notably, both models rank the features differently, indicating variations in their decision-making processes. For RF, the importance is predominantly given to speed-related features, such as maximum and standard deviation speed (V3 and V4). However, the Neural Network model tends to consider a more diverse set of features, encompassing speed-related attributes, GIS-related features (e.g., buffer ratios V10, V11), and even weather and socio-economic variables. This suggests that NN may exploit a broader range of information when making predictions compared to RF. The RBO analysis, as presented in Figure 4 (c & d), demonstrates a noticeable disparity in the similarity of feature importance rankings between RF and NN across different scenarios. For instance, in Scenario 1 for the imbalanced data, the similarity coefficient is moderately high at 0.87. However, as we move towards scenarios with lower GPS frequency, such as Scenario 6, the RBO value decreases substantially, reaching 0.68. This decline in similarity suggests that the feature importance rankings become increasingly dissimilar between RF and NN as the scenarios exhibit reduced GPS data precision. The trend is similar in the case of imbalanced data as well. However, when we observe the models' behavior in the sampled data, both RF and NN models assign importance to weather and socio-economic variables, indicating their potential significance in mode detection in data scenarios



Fig. 4: **Results of Rank-biased Overlap Analysis**

Table 3: **Accuracy Performance of Machine Learning Models Across Scenarios**

Transportation Mode	Scenario 1				Scenario 2				Scenario 3			
	RF	NN	KNN	SVM	RF	NN	KNN	SVM	RF	NN	KNN	SVM
Car	0.93	0.92	0.83	0.80	0.92	0.91	0.84	0.83	0.93	0.92	0.85	0.85
Train	0.69	0.35	0.00	0.20	0.59	0.51	0.00	0.10	0.62	0.51	0.00	0.00
Walk	0.93	0.91	0.71	0.70	0.90	0.86	0.70	0.68	0.86	0.86	0.73	0.72
Bicycle	0.87	0.78	0.18	0.17	0.87	0.83	0.29	0.29	0.83	0.80	0.29	0.29
Motorcycle	0.47	0.39	0.04	0.10	0.44	0.54	0.07	0.10	0.46	0.49	0.06	0.00
Bus	0.61	0.52	0.20	0.21	0.49	0.44	0.09	0.20	0.53	0.51	0.13	0.10
Overall	**0.90**	*0.86*	0.72	0.70	**0.87**	*0.85*	0.73	0.71	**0.87**	*0.86*	0.74	0.71
Transportation Mode	Scenario 4				Scenario 5				Scenario 6			
	RF	NN	KNN	SVM	RF	NN	KNN	SVM	RF	NN	KNN	SVM
Car	0.93	0.91	0.86	0.86	0.94	0.92	0.87	0.87	0.94	0.91	0.89	0.89
Train	0.75	0.44	0.04	0.00	0.60	0.37	0.00	0.00	0.57	0.31	0.00	0.00
Walk	0.90	0.89	0.80	0.80	0.93	0.91	0.76	0.77	0.93	0.91	0.70	0.77
Bicycle	0.85	0.79	0.37	0.20	0.91	0.86	0.52	0.3	0.82	0.74	0.51	0.30
Motorcycle	0.49	0.30	0.13	0.00	0.43	0.44	0.12	10	0.38	0.50	0.40	0.10
Bus	0.51	0.49	0.19	0.10	0.57	0.50	0.25	0.10	0.63	0.45	0.44	0.10
Overall	**0.89**	*0.85*	0.77	0.74	**0.89**	*0.86*	0.78	0.75	**0.88**	*0.84*	0.79	0.75

Table 4: **Class Balancing and Model Accuracy Performance Across Scenarios**

Transportation Mode	Original Imbalanced Data											
	Scenario 1		Scenario 2		Scenario 3		Scenario 4		Scenario 5		Scenario 6	
	RF	NN	RF	NN	RF	NN	RF	NN	RF	NN	RF	NN
Car	0.93	0.92	0.92	0.91	0.93	0.92	0.93	0.91	0.94	0.92	0.94	0.91
Train	0.69	0.35	0.59	0.51	0.62	0.51	0.75	0.44	0.6	0.37	0.57	0.31
Walk	0.93	0.91	0.9	0.86	0.86	0.86	0.9	0.89	0.93	0.91	0.93	0.91
Bicycle	0.87	0.78	0.87	0.83	0.83	0.8	0.85	0.79	0.91	0.86	0.82	0.74
Motorcycle	0.47	0.39	0.44	0.54	0.46	0.49	0.49	0.30	0.43	0.44	0.38	0.5
Bus	0.61	0.52	0.49	0.44	0.53	0.51	0.51	0.49	0.57	0.5	0.63	0.45
Overall	0.90	0.86	0.87	0.85	0.87	0.86	0.89	0.85	0.89	0.86	0.88	0.84
Transportation Mode	Sampled Balanced Data											
	RF	NN	RF	NN	RF	NN	RF	NN	RF	NN	RF	NN
Car	0.74	0.48	0.74	0.48	0.75	0.49	0.61	0.43	0.60	0.43	0.61	0.32
Train	0.80	0.55	0.75	0.52	0.74	0.52	0.79	0.50	0.74	0.5	0.67	0.46
Walk	0.92	0.89	0.86	0.79	0.90	0.82	0.85	0.86	0.8	0.8	0.75	0.80
Bicycle	0.89	0.8	0.83	0.79	0.83	0.79	0.81	0.78	0.75	0.65	0.66	0.63
Motorcycle	0.79	0.57	0.74	0.57	0.73	0.56	0.72	0.57	0.75	0.56	0.65	0.55
Bus	0.72	0.63	0.70	0.55	0.7	0.53	0.68	0.53	0.70	0.38	0.62	0.30
Overall	0.81	0.64	0.78	0.61	0.76	0.59	0.73	0.57	0.70	0.53	0.68	0.51

with varying class sizes.

Upon investigating the interpretability across different scenarios, we find that the behavior of both RF and NN models remains consistent. Although the same variables are generally ranked among the top five, the specific ranking of features varies across scenarios. This suggests that the models adapt their decision-making process based on changes in the data distribution, and certain features become more or less influential depending on the scenario. It was observed that as the frequency of GPS points reduced, buffer ratio for train stations (V10) and buffer ratio for bus station (V11) became the most important features (across RF and NN models for scenarios 4-6 in imbalanced

Table 5: **Ranking of Top Features based on SHAP Values**

RF Original					
Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
V3	V4	V4	V10	V10	V10
V1	V3	V3	V4	V1	V1
V4	V1	V1	V3	V3	V3
V10	V10	V10	V1	V4	V4
V7	V5	V7	V7	V7	V7
RF Sampled					
Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
V10	V10	V10	V10	V10	V1
V1	V4	V4	V4	V4	V4
V3	V3	V3	V3	V3	V7
V4	V1	V20	V1	V1	V3
V20	V7	V1	V20	V20	V6
NN Original					
Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
V3	V3	V3	V10	V11	V10
V1	V5	V4	V8	V10	V13
V4	V4	V1	V13	V3	V11
V7	V1	V7	V14	V4	V12
V5	V6	V5	V3	V13	V15
NN Sampled					
Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
V1	V1	V1	V6	V6	V6
V3	V10	V4	V1	V11	V10
V10	V11	V7	V10	V1	V11
V11	V20	V6	V4	V10	V1
V20	V3	V10	V11	V4	V16

data). This indicates how with reduced sampling of GPS points, the model relies less on variables derived from trajectory data and relies more on other variables. The observations from the RBO analysis in Figure 4 (b) reinforce these trends.

Moreover, when investigating the effects of class balancing, it was observed that RF models produce more similar feature importance ranking when compared between data types than NN models (see Figure 4 and Table 5). However, certain new variables become more important in the sampled datatset, they include, gender (V20, in both RF and NN) and temperature (in NN), which feature in top 5 ranking list in the sampled data, while they were not present in the imbalanced data.

The findings suggest that the model's understanding of the transportation mode classification task is heavily influenced by the context of data availability. Model type, GPS data frequency, and class-balancing seems to have an influence on what variables are the most important for transport mode detection. For effective implementation in practical settings, it becomes crucial to adapt feature engineering and model types based on the available data type. Often accuracy is made the sole indicator to make such decisions but aligning the model's interpretation with the context of data availability, more informed and reliable transportation mode detection can be achieved, supporting better decision-making in transportation planning and management.

5. Conclusion

This study aimed to investigate the influence of varying GPS trajectory frequencies and class balancing on travel mode detection using machine learning models. We conducted a comparative analysis of four algorithms—RF, NN, K-NN, and SVM—in six different scenarios, considering both class-imbalanced

and balanced datasets. The research provided valuable insights into the performance and interpretability of these models in transportation mode detection.

The key findings of our study can be summarized as follows: 1) RF models consistently outperformed other models in terms of accuracy, closely followed by NN models, with better performance than K-NN and SVM models. 2) In the imbalanced dataset, the detection accuracy for minority modes, such as train, bus, and bicycle, remained consistently low, even when the overall accuracy was high across scenarios with varying GPS data frequency. After class resampling to create a balanced dataset, the detection accuracy of previously minority modes substantially improved. Additionally, a decrease in GPS frequency led to reduced detection accuracy for both RF and NN models, establishing a clear link between class imbalance and GPS data frequency.3) The analysis of model interpretability demonstrated notable differences between model types, data types, and scenarios.

These research findings have significant policy implications for two stakeholder groups. Firstly, for designers of sensor-based travel survey applications, the results emphasize that both model accuracy and interpretability are significantly impacted by varying GPS data frequency. While the influence was less evident in the imbalanced dataset, the balanced data showed a near-linear relationship with decreasing GPS data frequency. However, interpretability analysis indicated that variables apart from those derived from trajectory data become more important in detection. Hence, the decision on how to balance this trade-off rests with the developers. Secondly, for analysts and policymakers tasked with choosing algorithms for mode detection, this study highlights the consistently good performance of RF models. Additionally, the interpretability analysis revealed that NN models utilize a broader range of variables compared to RF, making the former less intuitive. Similar disparities between model types have been observed in previous research (Chikaraishi et al., 2020).In conclusion, this study underscores the need for context-aware models capable of adapting to varying data frequency and feature importance. By aligning model interpretability with data availability, transportation planners and decision-makers can develop more informed and reliable transportation mode detection systems, supporting better decision-making in transportation planning and management.

In future research, other algorithms, including deep learning methods, could be further explored to evaluate their performance concerning GPS data frequency. Additionally, this study focused on the ex-post removal of GPS points from already collected trajectory data, and did not directly assess the impact of sampling incorporated within an application on accuracy and interpretability, presenting a notable limitation.

ACKNOWLEDGMENT: This work was done with the support of Japan Society for the Promotion of Science's (JSPS) KAKENHI grant (Grant no. 22K14339). The data used in this study was collected with the support of a grant from the Committee on Advanced Road Technology under the authority of the Ministry of Land, Infrastructure, Transport, and Tourism of Japan.

References

Abd Elrahman, S. M., & Abraham, A. (2013). A review of class imbalance problem. Journal of Network and Innovative Computing, 1(2013), 332–340.

Axhausen, K. W., Schönfelder, S., Wolf, J., Oliveira, M., & Samaga, U. (2003). 80 weeks of gps-traces: Approaches to enriching the trip information. Arbeitsberichte Verkehrs-Und Raumplanung(178).

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

Byon, Y. J., Ha, J. S., Cho, C. S., Kim, T. Y., & Yeun, C. Y. (2017). Real-time transportation mode identification using artificial neural networks enhanced with mode availability layers: A case study in dubai. Applied Sciences, 7(9), 923.

Chen, H., & Cheng, Y. (2023). Travel mode choice prediction using imbalanced machine learning. IEEE Transactions on Intelligent Transportation Systems, 24(4), 3795–3808.

Cheng, L., Chen, X., De Vos, J., Lai, X., & Witlox, F. (2019). Applying a random forest method approach to model travel mode choice behavior. Travel Behaviour and Society, 14, 1–10. doi: 10.1016/j.tbs.2018.09.002

Cottrill, C. D., Pereira, F. C., Zhao, F., Dias, I. F., Lim, H.-B., Ben-Akiva, M. E., & Zegras, P. C. (2013). Future mobility survey: Experience in developing a smartphone-based travel survey in singapore. Transportation Research Record, 2354, 59–67.

Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

Deniau, J. (2015). haversine. `https://github.com/mapado/haversine`.

Fix, E., & Hodges, J. L. (1951). Discriminatory analysis. nonparametric discrimination; consistency propert report number 4, project number 21-49-004, usaf school of aviation medicine, randolph field, tex (Tech. Rep.). (Reprinted as pp 261-279 of Agrawala, 1977)

Hashemi, S. R., Salehi, S. S. M., Erdogmus, D., Prabhu, S. P., Warfield, S. K., & Gholipour, A. (2018). Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. IEEE Access, 7, 1721–1735.

Ho, T. K. (1995). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278–282).

Jahangiri, A., & Rakha, H. A. (2015). Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. IEEE Transactions on Intelligent Transportation Systems, 16(5), 2406–2417.

Li, J., Miyashita, K., Kato, T., & Miyazaki, S. (2000). Gps time series modeling by autoregressive moving average method: Application to the crustal deformation in central japan. Earth, planets and space, 52, 155–162.

Li, J., Pei, X., Wang, X., Yao, D., Zhang, Y., & Yue, Y. (2021). Transportation mode identification with gps trajectory data and gis information. Tsinghua Science and Technology, 26(4), 403–416.

Lin, K., Kansal, A., Lymberopoulos, D., & Zhao, F. (2010). Energy-accuracy trade-off for continuous mobile device location. In Proceedings of the 8th international conference on mobile systems, applications, and services (pp. 285–298).

Lin, M., & Hsu, W.-J. (2014). Mining gps data for mobility patterns: A survey. Pervasive and Mobile Computing, 12, 1–16.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in neural information processing systems (pp. 4765–4774).

Marra, A. D., Becker, H., Axhausen, K. W., & Corman, F. (2019). Developing a passive gps tracking system to study long-term travel behavior. Transportation Research Part C: Emerging Technologies, 104, 348–368.

Maruyama, T., Sato, Y., Nohara, K., & Imura, S. (2015). Increasing smartphone-based travel survey participants. Transportation Research Procedia, 11, 280–288.

Nawarathne, K., Ball, R., Pereira, F. C., Zhao, F., Rahardjo, A. D., Zegras, C. P., & Ben-Akiva, M. E. (2014). Trade-off between smartphone battery life and stop detection accuracy. In Proc. of the 2014 ieee conf. on intelligent transportation systems (pp. 725–732).

Poucin, G., Farooq, B., & Patterson, Z. (2018). Activity patterns mining in wi-fi access point logs. Computers, Environment and Urban Systems, 67, 55–67.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386.

Statistics Bureau, M. o. I. A., & Communications. (2021). Population census of japan. https://www.stat.go.jp/english/data/kokusei/index.html.

Susan, S., & Kumar, A. (2021). The balancing trick: Optimized sampling of imbalanced datasets—a brief survey of the recent state of the art. Engineering Reports, 3(4), e12298.

Tamim Kashifi, M., Jamal, A. K. M., Samim Kashefi, M., Almoshaogeh, M., & Masiur Rahman, S. M. (2022). Predicting the travel mode choice with interpretable machine learning techniques: A comparative study. Travel Behaviour and Society, 29, 279–296.

Teranishi, Y., Kimata, T., Kawai, E., & Harai, H. (2019). Hybrid cellular-dtn for vehicle volume data collection in rural areas. In 2019 ieee 43rd annual computer software and applications conference (compsac) (Vol. 1, pp. 276–284).

Varghese, V. J., Chikaraishi, M., Urata, J., et al. (2020). Deep learning in transport studies: A meta-analysis on the prediction accuracy. Journal of Big Data Analytics in Transportation, 2(3), 199–220.

Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. ACM Transactions on Information Systems (TOIS), 28(4), 1–38.

Wessel, N., & Farber, S. (2019). On the accuracy of schedule-based gtfs for measuring accessibility. Journal of Transport and Land Use, 12(1), 475–500.

Wolf, J., Loechl, M., Myers, J., & Arce, C. (2001). Trip rate analysis in gps-enhanced personal travel surveys. In International conference on transport survey quality and innovation.

Xiao, G., Juan, Z., & Zhang, C. (2016). Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. Transportation Research Part C: Emerging Technologies, 71, 447–463.

(Received ??? ?? 202?)
(Accepted ??? ?? 202?)