Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 9 No. 1 (2024) 40 - 50 e-ISSN: 2580-0760

Prediction of Main Transportation Modes using Passive Mobile Positioning Data (Passive MPD)

Muhammad Farhan[1], Lya Hulliyyatus Suadaa[2]*, Sugiri[3], Alfatihah Reno MNSP Munaf[4], Setia Pramana[5]
[1,2,5]Politeknik Statistika STIS, Jakarta, Indonesia
[3,4]BPS Statistics Indonesia, Jakarta, Indonesia
[1]222011780@stis.ac.id, [2]lya@stis.ac.id, [3]sugiri@bps.go.id, [4]alfa@bps.go.id, [5]setia.pramana@stis.ac.id

Abstract

Indicators of the main mode of transportation used by domestic tourists during tourism trips cannot yet be estimated using Passive MPD which is recorded based on the location of the BTS that captures the cellular activity of domestic tourists. Previous research on identifying transportation modes from Passive MPD has its own shortcomings because it only relies on speed and travel time features. Meanwhile, there is Active MPD which is recorded using active geo-positioning and real-time, where the research involves many features and has a data structure similar to Passive MPD. Therefore, this research aims to conduct a study of the implementation of the method used to identify modes of transportation in Active MPDs to Passive MPDs as an approach to predicting the main modes of transportation. As a result, the transportation mode identification method in the Active MPD can be implemented in the Passive MPD. The best accuracy of 83.56% was obtained by the LightGBM model using all features. However, the Multinomial Logistic Regression model, which only uses 10 selected features, is the most effective and efficient model with an accuracy of 76.43% and a much shorter execution time.

Keywords: *prediction; Active MPD; Passive MPD; main transportation mode*

How to Cite: [. Farhan, L. H. Suadaa, Sugiri, A. R. M. N. S. P. Munaf, and S. Pramana, "Prediction of Main Transportation Modes using Passive Mobile Positioning Data (Passive MPD)", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 9, no. 1, pp. 40 - 50, Jan. 2025.
DOI: https://doi.org/10.29207/resti.v9i1.6128

1. Introduction

In terms of collecting data on domestic tourists, since 2018, BPS - Statistics Indonesia, has carried out data collection activities on domestic tourists in 34 provinces in Indonesia, which is then known as the Survei Wisatawan Nusantara (Indonesian domestic tourist survey) [1]. This survey activity was carried out conventionally and continued until 2019 [2]. In 2020, Survei Wisatawan Nusantara was conducted using a new method, namely Mobile Positioning Data (MPD), to overcome the weaknesses of conventional surveys which are only able to estimate up to the provincial level and are very dependent on respondent's memory, thus potentially causing errors in terms of respondent's answers, in addition to the potential for errors in survey sampling [3].

MPD is a large-scale dataset of transaction records and locations of customers from cellular operators (Mobile Network Operators / MNO) which are processed and stored in a system [3]-[5]. Based on the type of data collection, MPD used in geographic studies can be divided into Passive MPD and Active MPD [6], [7]. Passive MPD is location data that is stored automatically by the service provider system whenever a person's mobile phone interacts with the cellular network, such as call activity, sending or receiving messages, or internet access [3]-[5]. Data included in Passive MPD are Call Detail Record (CDR) and Location Based Advertising/Signalling (LBA/LBS) [3]-[5]. Meanwhile, Active MPD is tracking data for the location of a mobile phone that is determined using certain waves such as the Global Positioning System (GPS) [3]-[5], where GPS itself records the subscriber's position using active geo-positioning (representing the subscriber's actual location) and in real-time [8]-[11]. The use of MPD in collecting mobility data like domestic tourist statistics has several advantages [12], namely: Mobile phone use is widespread and popular in both developed and developing countries; The tendency of people to always carry mobile phones and make them important items; The initial data is in digital form so that

Received: 22-10-2024 | Accepted: 04-01-2025 | Published Online: 24-01-2025

it is free from human errors such as limited memory of respondents or data entry errors; and The use of MPD makes it possible to study the population movement in space and time dimensions more precisely.

Apart from the various advantages and potential benefits it has, MPD also has limitations that need to be taken into account, one of these limitations is that information regarding the characteristics of subscribers (in this case, domestic tourists) cannot be obtained if only use MPD without conducting a survey [13]. In the publication of 2020-2022 Statistik Wisatawan Nusantara (Indonesian domestic tourist statistics), MPD is only used to estimate the number of tourism trips and the average length of stay for domestic tourists [3]-[5].

Therefore, to obtain demographic characteristics (such as gender and age of tourists), travel patterns (such as the main purpose of the trip, types of tourist activities carried out, main modes of transport used, accommodation services used, average travel time) and average expenditure per trip by domestic tourists while travelling, BPS completes it with a Survei Digital Wisatawan Nusantara (digital survey of Indonesian domestic tourists) [3]-[5]. The use of digital surveys in collecting data on domestic tourists also has weaknesses similar to conventional surveys, namely that they are very dependent on the memory and response rate of respondents. Meanwhile, according to BPS [3]-[5] the response rate of survey respondents continues to decline and is followed by an increase in survey rejection by respondents.

These problems encourage further research regarding the use of MPD to estimate indicators that so far cannot be estimated using MPD. Of the several indicators mentioned previously, one of the indicators that is feasible to estimate using MPD is the indicator of the main mode of transportation used by domestic tourists during tourism trips. This is because there have been several previous studies that focused on identifying transportation modes from time-series sensor data such as GPS (Active MPD) [8]-[11] and CDR (Passive MPD) [14], [15].

Research by Kyaing et al. [14] uses a speed feature approach calculated from subscriber CDR data. Kyaing et al. [14] assumes that if a subscriber has a speed that is within a certain range, then it is believed that the subscriber is using a certain mode of transportation. However, according to Wang et al. [15], the application of a transportation mode identification method that only uses speed features has weaknesses in distinguishing transportation modes that have similar speeds, such as cars and buses. Therefore, Wang et al. [15] uses another feature approach in the form of travel time in identifying transportation modes. Wang et al. [15] believes that if the time spent by subscribers while traveling calculated from CDR data is close to the estimated travel time from Google Maps for certain modes of transportation, then the subscriber is believed to be using that mode of transportation.

However, the method applied in both studies using CDR data (Passive MPD) has its own shortcomings because it only relies on basic features such as speed and travel time. Passive MPD has low spatial accuracy, irregular spatial intervals, and quite long time gaps between records (can be minutes, hours, days, or even months) [16] so calculating speed and travel time features is considered inaccurate and does not represent actual conditions. Passive MPD has low spatial accuracy because the latitude and longitude coordinates in the Passive MPD record do not represent the actual location of the subscriber, but refer to the location of the Base Transceiver Station (BTS) which captures the subscriber's cellular transaction activity [3]-[5], [16]. On the other hand, Passive MPD has irregular spatial intervals and quite long time gaps between records because Passive MPD is only recorded when there is cellular transaction activity such as telephone transactions, short messages (SMS), internet access, etc. [3]-[5], [16].

Therefore, other methods are needed to identify modes of transportation that can overcome the weaknesses of Passive MPD. The solution that can be offered is to implement the method used to identify modes of transportation in the Active MPD to the Passive MPD. This is because apart from using features that involve speed and travel time in their calculations, research by Zheng et al. [8], Zhu et al. [9], Li et al. [10], and Witayangkurn et al. [11] also use other features calculated from GPS records (Active MPD) such as the total distance of the trip [8], [10], [11], heading change rate [8], [9], straight rate [10], percentage points that are on highways [11], percentage of points that are on railway lines [11], etc.

The transportation mode identification method in Active MPD is possible to be applied to Passive MPD because basically Active MPD has a data structure that is similar to Passive MPD, where one line/record of Active MPD and Passive MPD consists of trajectory/user id, latitude, longitude, and timestamp/datetime which indicates events from the trajectory/user [3]-[5], [8]-[11], [16]. Therefore, this research aims to conduct a study of the implementation of the methods used to identify modes of transportation from Active MPD such as GPS, to Passive MPD such as CDR and LBA/LBS as an approach to predicting the main mode of transportation used by domestic tourists during tourism trip. The main objectives of this study are to explore Passive MPD which is used to identify the main modes of transport, build a classification model using Passive MPD to classify the main modes of transport, and evaluate the performance of the classification model that has been built.

2. Research Methods

The research method applied in this study is shown in Figure 1.

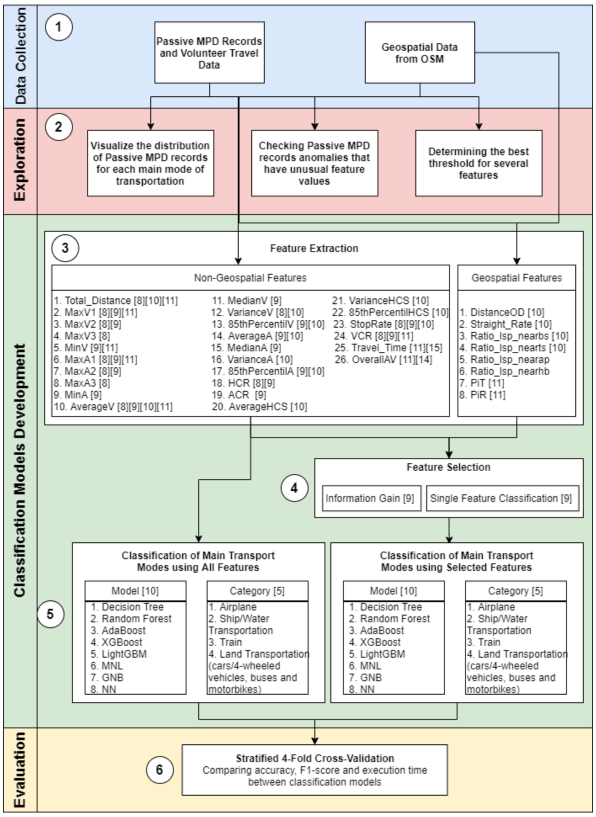

Figure 1. Research methods

2.1 Data Collection

In this research, the research method can be divided into four stages. The first stage is data collection consisting of Passive MPD records and geospatial data. To obtain Passive MPD records, this research collaborated with BPS as the government agency that have established cooperation with PT Telekomunikasi Selular (Telkomsel) and PT Indosat Tbk (Indosat Ooredoo Hutchison). Both telecommunications service companies provides access to their user's Passive MPD records.

Then, to get full access to Passive MPD records, volunteers from Telkomsel and Indosat users are needed, who are willing to provide access to their Passive MPD records through a written agreement regarding privacy data protection. This research also uses tourism trip data from volunteers which is obtained from the results of the internship program of Politeknik Statistika STIS for the 2022/2023 academic year using the interview method from January 23 to February 3, 2023. The tourism trip data is equipped with information regarding the main modes of transportation used by volunteers when travelling. Meanwhile, for geospatial data needs, this research uses Open Street Map (OSM) data because it has better precision and completeness than other geospatial data sources such as the Indonesian Earth Map (RBI) and Diva-GIS.

2.2 Exploration

After obtaining the data, the next step is to explore the Passive MPD records together with geospatial data. Exploration is carried out by looking at the distribution of Passive MPD records for each main transportation mode category, checking anomalies in the form of Passive MPD records which have unusual feature values, and determining the best threshold for several features whose calculations require a certain threshold value.

This exploration aims to determine the characteristics of Passive MPD records when used to identify the main mode of transportation. With this exploration, it is hoped that it can improve the accuracy of the main transportation mode classification results.

2.3. *Classification Models Development*

After the exploration of Passive MPD records is complete, feature extraction will be performed on each Passive MPD record that represents the volunteer's tourism trip. The extracted features can be classified into two categories, namely non-geospatial features and geospatial features. Non-geospatial features are dominated by indicators of speed, acceleration, and changes in movement direction which can be calculated directly from Passive MPD records. Meanwhile, to calculate geospatial features, it is necessary to involve geographic information obtained from OSM data in the previous stage.

After feature extraction is complete, the next stage is to select features using the Information Gain (IG) and Single Feature Classification (SFC). IG is a method in feature selection that uses a scoring technique to assign weights to a feature based on entropy values that have a maximum value [17]. Meanwhile, SFC is an approach in data processing and machine learning where only one feature or variable is used to perform classification or prediction. This feature selection process produces a set of features that play an important role in distinguishing the main transportation mode class categories. To assess the effectiveness of these features, in the next stage, a classification of the main modes of transportation is carried out using two scenarios. The first scenario is to classify the main transportation modes using all the features that have been extracted and the second scenario is to classify the main transportation modes using only the selected features resulting from feature selection.

Classification is carried out using supervised learning with several classification models as shown in Figure 1. Here is a brief explanation of the classification model used in this study.

Decision Tree (DT), the DT algorithm works by dividing the dataset into smaller subsets based on certain features, until a tree-like structure is formed, with each branch representing a decision or feature selection, and each leaf representing the final or class result [18].

Random Forest (RF), the RF algorithm uses a number of DTs to vote for the final classification result. When each DT is created, a part of the training samples is randomly selected. Not all the features are considered each time the split point is found when a tree splits. Only some candidate features chosen randomly are involved in finding the best split point [19].

Adaptive Boosting (AdaBoost) is a machine learning algorithm based on boosting techniques (i.e. models are trained sequentially, where each new model focuses on correcting the errors made by the previous model). This algorithm is designed to improve model accuracy by combining several simple models (usually very shallow decision trees or decision stumps) to form a more powerful final model [20].

Almost the same as AdaBoost, eXtreme Gradient Boosting (XGBoost) algorithm also uses boosting techniques. The difference is that XGBoost is a more advanced and optimized version of gradient boosting, with additional features such as regularization, parallel processing, and the ability to handle missing data [21].

Light Gradient Boosting Machine (LightGBM) is also a machine learning algorithm that uses boosting techniques, but is designed to overcome the limitations of traditional gradient boosting, especially in terms of processing speed and memory usage, especially on large datasets. Overall, LightGBM has higher speed and scalability than traditional boosting algorithms [22].

Multinomial Logistic Regression (MNL) is an extension of logistic regression used to model categorical dependent variables with more than two classes (multiclass). MNL predicts the probability of membership of each category in the dependent variable using the maximum likelihood estimation method based on several independent variables which can be categorical or continuous [23].

Gaussian Naïve Bayes (GNB) utilizes the Gaussian distribution to represent variables in the classification process. Each feature in the training data is assumed to follow a Gaussian distribution, making it easier to calculate probabilities and make decisions in the classification process [24].

Neural Network (NN), is a mathematical model inspired by the structure and function of human biological neural networks [25]. NN consists of information processing units called neurons or nodes, which are organized in layers. These layers involve the input layer, hidden layer, and output layer. Each connection between neurons has a weight that can be changed during the model training process [26].

The features extracted from each Passive MPD record (both in the first and second scenario) along with the main mode of transportation used by volunteers during their trip will be used as input variables for training and evaluating classification models. The main modes of transportation used in this research refer to the main modes of transportation in the Statistik Wisatawan Nusantara 2022 and Survei Digital Wisatawan Nusantara 2023 which are categorized based on the terrain or routes they traverse.

2.4 Evaluation

Finally, to evaluate the classification model that has been built using two different scenarios, the stratified k-fold cross-validation with the number k=4 is used. Stratified k-fold cross-validation works by dividing the dataset into k equal parts, where each part (fold) maintains the same proportion of the target/dependent variable (in the training and testing process) as in the original dataset [27].

This evaluation method can produce comparisons of accuracy, F1-score, and also the execution time between classification models both in the first scenario and the second scenario, as well as comparing which scenario is more effective and efficient in classifying the main modes of transportation according to their class categories.

3. Results and Discussions

3.1 Data Description

A total of 225 volunteer tourism trip data in the period 01 January 2022 – 30 November 2022 with travel area coverage throughout Indonesia along with 25900 Passive MPD records which represent these tourism trips were used in this research. Each Passive MPD record consists of subscriber ID (hashing), datetime, source, latitude, longitude, province ID, district ID, subdistrict ID, event month and each volunteer tourism trip data consists of subscriber ID (hashing), month of travel, week of travel, main mode of transportation, length of trip (in hours), province of origin, district of origin, subdistrict of origin, province of destination, district of destination, subdistrict of destination. These two data can be matched via subscriber ID to obtain Passive MPD records with the corresponding main mode of transportation. Of the 225 tourism trip data, there were 66 trips using airplanes as the main mode of transportation, 4 trips using ships/water transportation, 34 trips using trains, and 121 trips using land transportation such as motorbikes, cars/4-wheeled vehicles, or buses.

3.2 Data Exploration

Data exploration is carried out by looking at the distribution of Passive MPD records for each main transportation mode category, checking anomalies in the form of Passive MPD records which have unusual feature values, and determining the best threshold for several features whose calculations require a certain threshold value.

Passive MPD record distribution: The following Figure 2 - Figure 5 displays the distribution of Passive MPD records from one sample of tourism trips taken randomly for each main mode of transportation. Based on the visualization, it can be seen that the distribution of Passive MPD records from tourism trips for each main mode of transportation is quite different. This difference mainly occurs in the number of records/points that represent the tourism trip. From

Figure 2 - Figure 5, it can be identified that tourism trips that use airplanes as the main mode of transportation only have a few records/points on their journey that cover quite long distances. This also happens on tourism trips using ships/water transportation where there are no records recorded in the ocean area so this reduces the total number of records/points on the tourism trip.

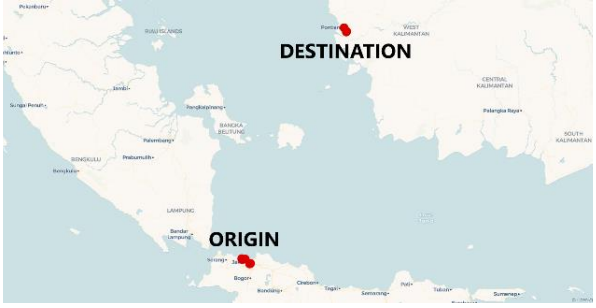

Figure 2. MSISDN00088 tourism trip from East Jakarta to Pontianak by airplane

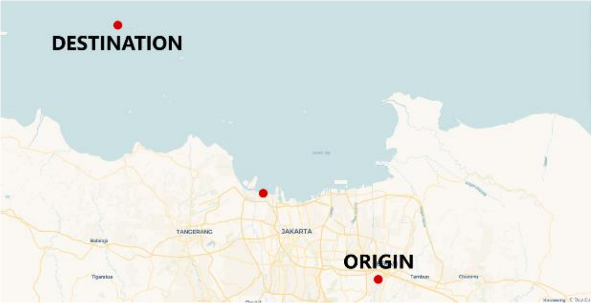

Figure 3. MSISDN00007 tourism trip from Bekasi to Kepulauan Seribu by ship/water transportation

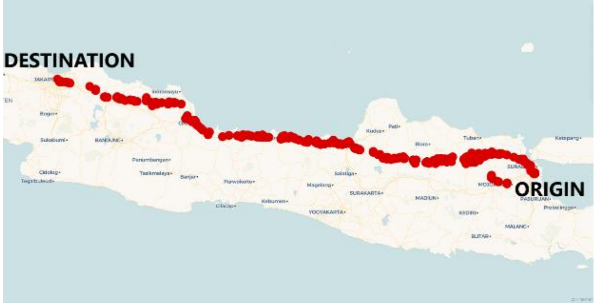

Figure 4. MSISDN00209 tourism trip from Surabaya to Jakarta Pusat by train

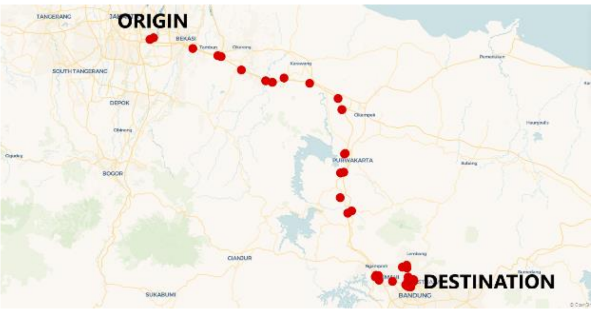

Figure 5. MSISDN00002 tourism trip from Jakarta Timur to Bandung Barat by land transportation

Meanwhile, on tourism trips that use trains and land transportation such as motorbikes, cars/4-wheeled vehicles, or buses as the main mode of transportation, it can be seen that there are many records/points along the trip that cover quite long distances. So that the distance traveled and the number of records/points seems balanced. This phenomenon encourages the need for a new calculation feature to calculate the number of Passive MPD records/points per unit distance travelled on each volunteer trip. It is hoped that this feature can differentiate the main mode of transportation used by volunteers during their trip. This feature can then be called PiTP (Point in Travel Period) and can be calculated using Formula 1.

$$\mathrm{PiTP} = \frac{count_{point}}{\mathrm{DistanceOD}} \tag{1}$$

$count_{point}$ represents the number of Passive MPD records/points along the trip and DistanceOD represents the haversine distance between the origin and destination points of a trip.

Anomalies in speed features: Of the total of 25900 Passive MPD records representing 225 tourism trips, there are 5239 records that have a speed of more than 1000 km/hour. This value is certainly unusual because according to Thinkmetric [28], the highest reasonable speed among the four main modes of transportation used in this research is 1000 km/hour, which is the maximum speed of an airplane. One of the reasons for the existence of records that have abnormal speeds is record lag, which is a condition when the time for recording the position of a mobile device does not match the actual time where the device is located. Examples of lag records can be seen in Table 1.

Apart from that, anomalies in the speed feature also occur when there is more than one record with different locations overlapping at the same time. In real cases, this is not possible because a subscriber cannot possibly be in two or more different locations at the same time. This event causes the speed feature calculation to return a NULL value. Examples of overlapping records can be seen in Table 1.

Table 1. Example of lag record and record overlapping

MSISDN	Datetime	Latitude	Longitude	Speed
msisdn42	2022-11-19 12:24:08	-6.90073	107.6283	0
msisdn42	2022-11-19 12:24:09	-6.90073	107.6283	0
msisdn42	2022-11-19 12:24:10	-6.89871	107.6235	2082.8[a]
msisdn42	2022-11-19 12:24:15	-6.89871	107.6235	0[b]
msisdn42	2022-11-19 12:24:15	-6.90073	107.6283	NULL[b]
msisdn42	2022-11-19 12:24:20	-6.90073	107.6283	0

[a]Example of a lag record that causes an unusual speed
[b]Examples of overlapping records cause the speed to be NULL

Lag records and overlapping records must be eliminated because they can cause bias in calculating features involving speed such as average speed, average acceleration, maximum speed, maximum acceleration, speed variance, acceleration variance, etc.

Anomalies in geospatial features: Anomalies occur in the geospatial features Ratio_lsp_nearts, Ratio_lsp_nearbs, Ratio_lsp_nearap, and Ratio_lsp_nearhb. Ratio low-speed point (Ratio_lsp) is calculated by rationing points that have a speed of less than equal to 1 m/s (3.6 km/hour) within a distance of 500 meters from the closest train station (nearts), bus station (nearbs), airport (nearap) or harbor (nearhb) to all points that have a speed of fewer than 1 m/s for each tourism trip. In the research of Li et al. [12], if no point has a speed less than 1 m/s, then the values of Ratio_lsp_nearts, Ratio_lsp_nearbs, Ratio_lsp_nearap, and Ratio_lsp_nearhb are set as -1.

However, after calculating, there were 57 trips that had Ratio_lsp_nearts, Ratio_lsp_nearbs, Ratio_lsp_nearap, and Ratio_lsp_nearhb values equal to -1. This shows that there are quite a lot of tourism trips that do not have points/records with speeds less than 1 m/s and this phenomenon can certainly reduce the richness of Passive MPD records because if there are points/records that are within 500 meters of the closest train station, bus station, airport or harbor but with a speed of more than 1 m/s, then that point will not be counted.

This problem encourages the need to calculate new features without involving a speed threshold so that all points/records can be involved in the calculation. For this reason, in this study, a new geospatial feature was added in the form of the percentage of points that are within 500 meters of the nearest train station (Pi_nearts), bus station (Pi_nearbs), airport (Pi_nearap) or harbor (Pi_nearhb) to all points in each trip. This new feature is expected to provide significant information in distinguishing the main modes of transportation used by volunteers during tourism trips.

Determining the best threshold on several features: Some features whose calculations require a certain threshold are HCR, ACR, StopRate, and VCR. To get the best threshold, Zheng et al. [8] and Zhu et al. [9] applied Single Feature Classification (SFC) in their research, namely classifying the main modes of transportation using only one feature with different threshold values and then comparing the accuracy results with each other.

In this research, the SFC process for selecting the best threshold is carried out using the Random Forest (RF) model. Figure 6 until Figure 9 displayed the average accuracy of the main transportation mode classification results for each feature with the threshold values tested referring to Zheng et al. [8] for HCR, StopRate, and VCR and Zhu et al. [9] for ACR.

Based on Figure 6 to Figure 9, the results show that the best threshold for HCR=39 degrees, StopRate=1 m/s (3.6 km/h), VCR=4.6 m/s (16.56 km/h), and ACR=0.16 m/s^2 (2073.6 km/h^2) because it has the highest average accuracy of the main transportation mode classification results compared to other thresholds for each feature. This best threshold will later be used for actual feature calculations in the modelling process.

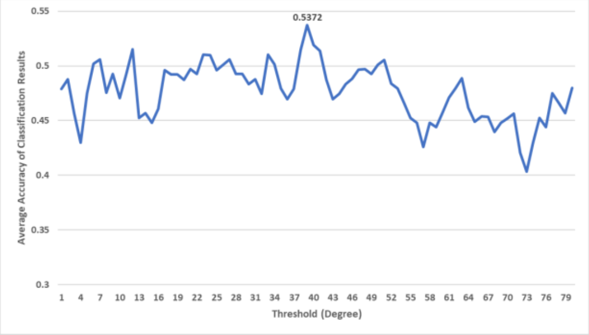

Figure 6. Selection of the best threshold for HCR

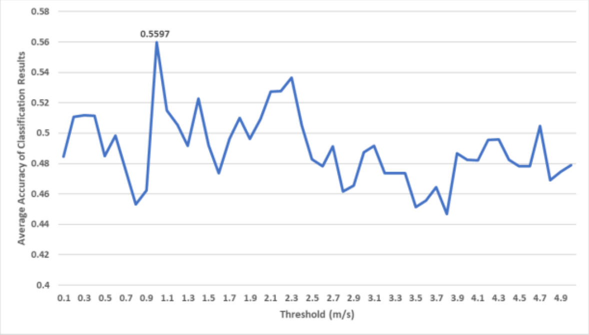

Figure 7. Selection of the best threshold for StopRate

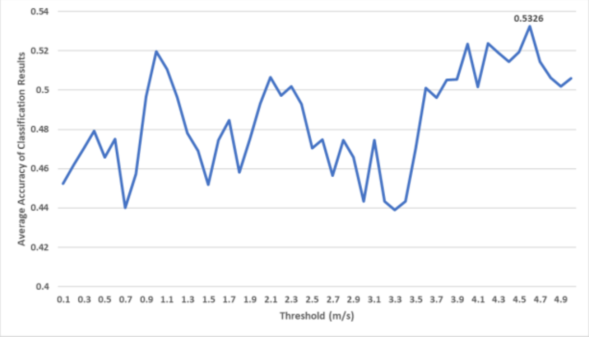

Figure 8. Selection of the best threshold for VCR

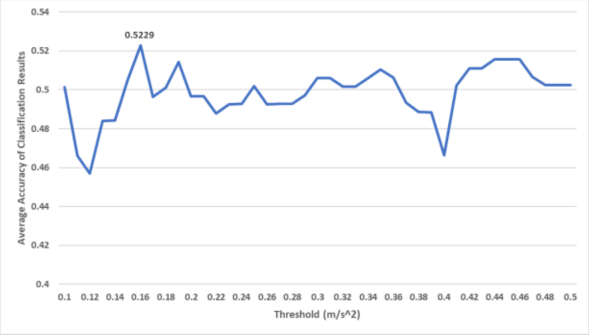

Figure 9. Selection of the best threshold for ACR

3.3 Classification Models Development

The process of building a classification model consists of three stages, namely feature extraction, feature selection, and classification with two scenarios. The first scenario is to classify the main transportation modes using all the features that have been extracted and the second scenario is to classify the main transportation modes using only the selected features resulting from feature selection. All stages in the classification model development process (including evaluation) were carried out using Google Colaboratory with a CPU and RAM of 12.7 GB.

Table 2. The average value of several features for each main mode of transportation

Feature	Main Mode of Transportation			
	Airplane	Ship/ Water Transportation	Train	Land Transportation
Total Distance (km)	904.04	320.55	505.00	212.71
AverageV (km/h)	126.34	75.45	130.24	98.72
Travel Time (h)	14.72	15.22	17.44	13.56
PiTP	0.16	0.24	0.40	0.98
Pi nearbs (%)	13.32	3.44	4.51	5.17
Pi nearts (%)	8.82	5.38	11.39	7.44
Pi nearap (%)	42.33	0.00	0.49	0.13
Pi nearhb (%)	0.70	17.58	0.00	0.00

Feature extraction is carried out for each Passive MPD which represents a volunteer tourism trip. From this feature extraction process, the characteristics of each main mode of transport used by volunteers during their trip can be identified by calculating the average feature value for each main mode of transport. In Table 2, several features are presented with their average values for each main mode of transportation.

From Table 2, the average speed (AverageV) of airplanes (126.34 km/hour) is slower than the average speed of trains (130.24 km/hour). This is in contrast to the situation in the real world, where the speed of a plane should be much faster than the speed of a train. This problem indicates that the calculation of speed features from Passive MPD records is not accurate enough and does not represent actual conditions. This result further reinforces the statement about the weaknesses of the speed feature as explained in the background.

Feature selection is carried out using the Information Gain (IG) and Single Feature Classification (SFC) methods by applying several schemes, including the following.

Each classification model will perform modelling based on the features selected using IG, with the number of features chosen to be 5, 10, and 15 features.

Each classification model will perform modelling based on the features selected using SFC, with the number of features chosen to be 5, 10, and 15 features.

Each classification model will perform modelling using the same/overlapping features resulting from feature selection using IG and SFC, with the number of selected features being 5, 10, and 15.

The model used in SFC in points 2 and 3 is adjusted to match the classification model used for modelling. From these three schemes, a total of 72 feature selection combinations will be modelled in the second scenario.

Main Transportation Mode Classification: After feature extraction and selection are complete, the next stage is to classify the main modes of transportation using the Decision Tree (DT), Random Forest (RF), Adaptive Boosting (AdaBoost), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Multinomial Logistic Regression (MNL), Gaussian Naïve Bayes (GNB), and Neural Network (NN) for two scenarios, namely using all features (first scenario) and using selected features resulting from feature selection (second scenario).

In modelling, this research uses the best parameters for each classification model (except MNL, GNB, and NN) which are obtained by comparing the average accuracy of classification results from the stratified cross-validation process with a total of k=4 for each previously determined parameter combination. In the stratified cross-validation process, for each iteration, the training data is first resampled using the Synthetic Minority Over-sampling Technique (SMOTE) method. This aims to balance the amount of data in the main transportation mode class categories so that the model does not tend to predict the majority class and ignore the minority class.

Then, for the testing scenario, the accuracy and F1-score of each classification model will be calculated for the first scenario (using all features) and the second scenario (using selected features obtained from feature selection). The second scenario allows each classification method to have 9 models. For example, in the case of the Decision Tree (DT), there are 9 model combinations in the second scenario, including 3 DT models with features selected using the Information Gain, where each model has 5 features, 10 features, and 15 features, respectively. Then, 3 DT models with features were selected using the Single Feature Classification, where each model also has 5 features, 10 features, and 15 features, respectively. Lastly, 3 DT models with features selected from the intersection of features were obtained using both the Information Gain and Single Feature Classification, with each model having 5 features, 10 features, and 15 features, respectively.

3.4 *Classification Model Evaluation*

The classification model evaluation was conducted using stratified k-fold cross-validation with k=4. The choice of k=4 was made to ensure that the minority class, which is the main mode of ship/water transportation with only 4 samples, is evenly distributed across the validation data in each iteration (fold) of the cross-validation process. This ensures that no iteration (fold) ends up with validation data lacking samples of the main mode of ship/water transportation. A total of 80 classification models, consisting of 8 models in the first scenario (using all features) and 72 models in the second scenario (using selected features from feature selection), will be compared in terms of average accuracy, average F1 score, and execution time.

Table 3. The results of the overall modelling evaluation in the first scenario

Model	Overall Evaluation (%)		Execution Time (seconds)		
	Avg. F1-score	Avg. Accuracy	Feature Extraction	Model Training	Model Testing
DT	50.12	79.54	2206.9351	0.0059	0.0013
RF	54.62	80.46	2206.9351	0.1577	0.0065
Ada Boost	67.66	82.24	2206.9351	4.2903	0.0981
XG Boost	67.11	80.88	2206.9351	0.2716	0.0078
Light GBM	69.64	83.56	2206.9351	0.1964	0.0033
MNL	53.29	67.11	2206.9351	0.1936	0.0004
GNB	12.38	12.04	2206.9351	0.0060	0.0025
NN	13.34	40.83	2206.9351	0.4998	0.1389

Note: execution time carried out in Google Colaboratory using a CPU with 12.7 GB RAM

Based on Table 3, it can be seen that the LightGBM is the most effective and efficient model in classifying the main modes of transportation in the first scenario. Effective here means that the LightGBM model can properly classify the main modes of transportation into actual class categories. This is proven by the LightGBM which has the highest average accuracy and F1-score among other classification models, namely 83.56% for average accuracy and 69.64% for average F1-score.

While efficient here means that the LightGBM modelling process only requires or consumes a fairly short execution time. The LightGBM model only spends an average execution time of around 0.1964 seconds for the training process and 0.0033 seconds for the testing process per 225 tourism trip data, where the average execution time for the testing process is much shorter compared to several other classification models such as RF, AdaBoost, XGBoost and NN. As for the execution time spent in carrying out the feature extraction process, the eight classification models in the first scenario both use all the features, so evaluation of the execution time for the feature extraction process cannot be carried out.

Then, to see how the LightGBM model performs in classifying each main transportation mode category, it can be seen from the average F1 score per main transportation mode class category as shown in

Table 4.Based on Table 4, it can be seen that the LightGBM model is very good in classifying airplane (F1-score = 92.16%) and land transportation (F1-score = 87.64%), but not good enough in classifying ship/water transportation (F1-score = 50.00%) and train (F1-score = 48.74%).

Table 4. The result of LightGBM modelling evaluation per category of main transportation mode in the first scenario

Model	Main Mode of Transportation	Average F1-score (%)
Light GBM	Airplane	92.16
	Ship/Water Transportation	50.00
	Train	48.74
	Land Transportation	87.64

Next, to evaluate the modelling process in the second scenario, each classification method will select one model that is the most effective (in terms of average accuracy and F1-score) and efficient (in terms of average execution time). The best model from each classification method will then be compared with one another, and one model will be selected as the best at classifying the main transportation modes according to their class categories in the second scenario. The comparison of the best models between classification methods in the second scenario is shown in Table 5.

Based on Table 5, the Multinomial Logistic Regression (MNL) model with 10 features resulting from feature selection using SFC method is the most effective and efficient model in classifying the main modes of transportation in the second scenario. This model was chosen with several considerations as follows.

Even though this MNL model has the third highest average accuracy of classification results (76.43%), the average F1-score of this model is the highest among the other models (72.32%).

This MNL model also has the highest average precision and recall values compared to other models, namely 76.56% for average precision and 73.69% for average recall.

Lastly, the MNL model has a relatively short average execution time for feature extraction (61.6822 seconds), model training (0.1659 seconds), and model testing (0.0019 seconds) per 225 tourism trip data. This average execution time is much shorter compared to the model with the highest average accuracy, the LightGBM model (79.10%), which takes an average execution time of 196.5744 seconds for feature extraction, 0.4772 seconds for model training, and 0.0073 seconds for model testing per 225 tourism trip data.

Then, to find out what features play an important role in differentiating the main transportation mode class categories, Table 6 shows the 10 most frequently used features by the best model in the second scenario. Based on Table 6, the HCR, DistanceOD, Total_Distance, ACR and Ratio_lsp_nearts are the five features most frequently used for modelling. The HCR, DistanceOD and Total_Distance are used by the seven best models in the second scenario, while the ACR and Ratio_lsp_nearts are used by the six best models in the second scenario. This shows that these five features have an important role in differentiating the main modes of transportation according to their class categories.

Table 5. Overall comparison of the best models across classification methods in the second scenario

Model	Feature Selection Method	Number of Features	Overall Evaluation (%)				Execution Time (seconds)		
			Average Precision	Average Recall	Average F1-score	Average Accuracy	Feature Extraction	Model Training	Model Testing
DT	SFC	15	74.64	55.80	50.41	74.66	45.4967	0.0040	0.0013
RF	SFC	15	62.04	50.09	49.57	76.00	28.2658	0.2507	0.0096
AdaBoost	SFC	15	69.27	61.80	61.48	76.02	45.4967	2.2041	0.0978
XGBoost	SFC	10	67.94	68.31	64.19	76.88	45.1042	0.1676	0.0048
LightGBM	SFC	15	74.01	70.48	68.53	79.10	196.5744	0.4772	0.0073
MNL	SFC	10	76.56	73.69	72.32	76.43	61.6822	0.1659	0.0019
GNB	IG	5	54.89	55.62	48.55	59.99	24.4835	0.0028	0.0014
NN	IG & SFC (Overlapping Features)	5	47.15	26.56	16.06	42.61	8.1105	1.2041	0.1969

Note: execution time carried out in Google Colaboratory using a CPU with 12.7 GB RAM

Table 6. The ten most frequently used features of the best model from each classification method in the second scenario

Feature	DT	RF	AdaBoost	XGBoost	LightGBM	MNL	GNB	NN	Total
HCR	√	√	√	√	√	√		√	7
DistanceOD	√	√	√	√	√	√	√		7
Total_Distance	√	√	√	√	√	√	√		7
ACR	√	√	√	√	√	√			6
Ratio_lsp_nearts	√	√	√	√	√	√			6
VCR	√	√	√	√	√				5
StopRate	√	√	√		√	√			5
OverallAV	√	√	√		√			√	5
Pi_nearhb	√		√	√	√	√			5
85thPercentileHCS	√	√	√	√	√				5

Next, the best classification model in the second scenario, the MNL model, will be compared with the best classification model in the first scenario, the LightGBM model. Tables 7 and 8 present a comparison of the two best models from each scenario.

Table 7. Average F1-score and accuracy of the best model in each scenario

Scenario	Model	Number of Features	Average F1-score (%)	Average Accuracy (%)
I	LightGBM	39	69.64	83.56
II	MNL	10	72.32	76.43

Table 8. Execution time for the feature extraction, model training, and model testing process from the best model in each scenario

Scenario	Model	Number of Features	Execution Time (seconds)		
			Feature Extraction	Model Training	Model Testing
I	Light GBM	39	2206.9351	0.1964	0.0033
II	MNL	10	61.6822	0.1659	0.0019

Note: execution time carried out in Google Colaboratory using a CPU with 12.7 GB RAM

Based on Tables 7 and 8, it can be concluded that the MNL model with 10 features is the most effective and efficient model for classifying the main transportation modes according to their class categories. Although the MNL has a lower average accuracy (76.43%) compared to the LightGBM (83.56%), MNL has a higher average F1-score of 72.32% compared to the LightGBM, which only has an average F1-score of 69.64%.

In addition, the comparison of execution times which is an important part of modeling is also won by the MNL model. As shown in Table 8, the MNL model had a much shorter average execution time, taking only about 61.6822 seconds for the feature extraction process, 0.1659 seconds for the model training process, and 0.0019 seconds for the model testing process, compared to the LightGBM model, which took an average of up to 2206.9351 seconds for the feature extraction process, 0.1964 seconds for the model training process, and 0.0033 seconds for the model testing process per 225 tourism trip data. The significant difference in execution time, especially in the feature extraction process, is due to the number of features used in the modelling, with the MNL model using only 10 features, while the LightGBM model uses all 39 features.

Table 9. The evaluation results of the best model (MNL) per category of main transportation mode

Scenario	Model	Number of Features	Main Mode of Transportation	Average F1-score (%)
II	MNL	10	Airplane	89.23
			Ship/Water Transportation	75.00
			Train	45.17
			Land Transportation	79.89

Then, from the average F1-score per class category of main transportation modes for the MNL model, as shown in Table 9, it can be seen that the MNL model performs very well in classifying airplanes as the main transportation mode (F1-score = 89.23%), performs fairly well in classifying land transportation (F1-score = 79.89%) and ship/water transportation (F1-score = 75.00%), but does not perform well in classifying trains as the main transportation mode (F1-score = 45.17%).

4. Conclusions

Based on the research results, the method for identifying transportation modes in Active MPD can be implemented for Passive MPD. Various explorations were successfully conducted, including examining the distribution of Passive MPD records for each category of main transportation modes, checking for anomalies in Passive MPD records with unusual feature values, and determining the best thresholds for several features. These explorations revealed the need for calculating new features and the elimination of Passive MPD records with unusual feature values to improve classification accuracy. The process of building a classification model consists of three steps, including feature extraction, feature selection, and classification of the main transportation modes with two scenarios. The first scenario is to classify the main transportation modes using all the features resulting from the feature extraction process and the second scenario is to classify the main transportation modes using the selected features resulting from feature selection. A total of 80 classification models consisting of 8 models in the first scenario and 72 models in the second scenario were successfully produced in this process. Among all the classification models, the Multinomial Logistic Regression (MNL) model with 10 features selected through the Single Feature Classification (SFC) method is the most effective and efficient in classifying the main transportation modes according to their class categories. This model has an average accuracy of 76.43% and an average F1-score of 72.32%, with an average execution time for the feature extraction, model training, and model testing processes of approximately 61.6822 seconds, 0.1659 seconds, and 0.0019 seconds, respectively, per 225 tourism trip data. Based on the research that has been carried out, suggestions that can be applied in further research are to add more tourism trip data and related research references as an effort to increase the accuracy of the prediction results for the main modes of transportation used by domestic tourists during tourism trips.

References

[1] Badan Pusat Statistik, “Statistik Wisatawan Nusantara 2018,” Jakarta, 2019.

[2] Badan Pusat Statistik, “Statistik Wisatawan Nusantara 2019,” Jakarta, 2020.

[3] Badan Pusat Statistik, “Statistik Wisatawan Nusantara 2020,” Jakarta, 2021.

[4] Badan Pusat Statistik, “Statistik Wisatawan Nusantara 2021,” Jakarta, 2022.

[5] Badan Pusat Statistik, “Statistik Wisatawan Nusantara 2022,” Jakarta, 2023.

[6] R. Ahas, A. Aasa, S. Silm, and M. Tiru, *Mobile Positioning Data in Tourism Studies and Monitoring: Case Study in Tartu, Estonia*. 2007. doi: 10.1007/978-3-211-69566-1_12.

[7] A. Kuusik, M. Tiru, R. Ahas, and U. Varblane, “Innovation in destination marketing: The use of passive mobile positioning for the segmentation of repeat visitors in Estonia,” *Balt. J. Manag.*, vol. 6, pp. 378–399, Sep. 2011, doi: 10.1108/17465261111168000.

[8] Y. Zheng, Y. Chen, Q. Li, X. Xie, and W.-Y. Ma, “Understanding Transportation Modes Based on GPS Data for Web Applications,” *TWEB*, vol. 4, Jan. 2010, doi: 10.1145/1658373.1658374.

[9] Q. Zhu *et al.*, *Identifying Transportation Modes from Raw GPS Data:*, vol. 623. 2016. doi: 10.1007/978-981-10-2053-7_35.

[10] J. Li, X. Pei, X. Wang, D. Yao, Y. Zhang, and Y. Yue, “Transportation mode identification with GPS trajectory data and GIS information,” *Tsinghua Sci. Technol.*, vol. 26, pp. 403–416, Aug. 2021, doi: 10.26599/TST.2020.9010014.

[11] A. Witayangkurn, T. Horanont, N. Ono, Y. Sekimoto, and R. Shibasaki, *Trip Reconstruction and Transportation Mode Extraction on Low Data Rate GPS Data from Mobile Phone*. 2013.

[12] R. Ahas, S. Silm, O. Järv, E. Saluveer, and M. Tiru, “Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones,” *J. Urban Technol.*, vol. 17, pp. 3–27, Apr. 2010, doi: 10.1080/10630731003597306.

[13] R. Ahas, S. Silm, E. Saluveer, and O. Järv, “Modelling Home and Work Locations of Populations Using Passive Mobile Positioning Data,” in *Location based services and telecartography II: From sensor fusion to context models*, 2009, pp. 301–315. doi: 10.1007/978-3-540-87393-8_18.

[14] Kyaing, K. K. Lwin, and Y. Sekimoto, “Identification of various transport modes and rail transit behaviors from mobile CDR data: A case of Yangon City,” *Asian Transp. Stud.*, vol. 6, p. 100025, 2020, doi: https://doi.org/10.1016/j.eastsj.2020.100025.

[15] H. Wang, F. Calabrese, G. Di Lorenzo, and C. Ratti, “Transportation mode inference from anonymized and aggregated mobile phone call detail records,” in *13th International IEEE Conference on Intelligent Transportation Systems*, 2010, pp. 318–323. doi: 10.1109/ITSC.2010.5625188.

[16] A. P. Putra, W. O. Z. Madjida, I. A. Setyadi, A. R. S. Nugroho, and A. R. M. Munaf, “AMDA: Anchor Mobility Data Analytic for Determining Home-Work Location from Mobile Positioning Data ,” *Proc. Int. Conf. Data Sci. Off. Stat.*, vol. 2021, no. 1 SE-Data Science, pp. 296–304, Jan. 2022, doi: 10.34123/icdsos.v2021i1.239.

[17] S. Rokhmah and N. A. Rozaq Rais, “APPLICATION OF DATA MINING FOR PREDICTION OF LONG COVID ON COVID-19 SURVIVAL WITH FEATURE SELECTION AND NAÏVE BAYES METHOD,” *J. Tek. Inform.*, vol. 3, no. 5 SE-Articles, pp. 1397–1405, Oct. 2022, doi: 10.20884/1.jutif.2022.3.5.561.

[18] C. J. S. Leo Breiman, Jerome Friedman, R.A. Olshen, *Classification and Regression Trees*, 1st ed. New York: Chapman and Hall/CRC, 1984. doi: https://doi.org/10.1201/9781315139470.

[19] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[20] R. E. Schapire, “A brief introduction to boosting,” in *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, in IJCAI’99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 1401–1406.

[21] T. Chen and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*. 2016. doi: 10.1145/2939672.2939785.

[22] G. Ke *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[23] J. Starkweather and A. K. Moske, “Multinomial Logistic Regression,” Texas, 2011.

[24] R. Mubarak, M. Hanafi, and D. Sasongko, “Komparasi Performa Naive Bayes Gaussian dan K-NN Untuk Prediksi Kelulusan Mahasiswa dengan CRISP-DM,” *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 6, p. 2982 2991, 2024.

[25] I. Mekongga, R. Gernowo, and A. Sugiharto, “The Prediction of Bandwidth On Need Computer Network Through Artificial Neural Network Method of Backpropagation,” *J. Sist. Inf. Bisnis; Vol 2, No 2 Vol. 2 Nomor 2 Tahun 2012DO - 10.21456/vol2iss2pp098-107*, Jun. 2012, [Online]. Available: https://ejournal.undip.ac.id/index.php/jsinbis/article/view/40

[26] V. Y. P. Ardhana *et al.*, “Prediksi Flight Delay Berbasis Algoritma Neural Network,” *J. Informatics, Electr. Electron. Eng.*, vol. 2, no. 1, 2022.

[27] S.Suganya and N.Kamalraj, “A SURVEY ON CREDIT CARD FRAUD DETECTION,” 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:212444345

[28] “Thinkmetric. Speed.” [Online]. Available: https://thinkmetric.uk/basics/speed/