Transport Mode Detection in a Smartphone-based National Travel Survey

Research Report (Repurposed for Markup cCurse

Quinty Boer (4845242)

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Advisors:

Yvonne A.P.M. Gootzen (CBS, TU/e)

Dr. Jonas Klingwort (CBS)

Dr. Peter Lugtig (UU)

Daniëlle Remmerswaal (UU)

FETC-approval: 24-1967

Candidate Journal: Transportation Research Interdisciplinary Perspectives

Word count: 2494

1 Introduction

Understanding travel behaviour in a society is important for national infrastructure planning. Traditionally, information on travel behaviour is collected by national statistical institutes through diary surveys. In these types of surveys, respondents are asked to report in detail on all their trips over a period of time. Often, the requested information on travel behaviour includes the travel mode, time, duration, route, and purpose. Diary-based studies rely on the respondent's ability to remember their travel activities, which may be susceptible to under-reporting and response-over-time bias (McCool et al., 2021; Prelipcean et al., 2018). Furthermore, these studies impose a high burden on the respondents, which is associated with an increased risk of non-response and dropout (Wang et al., 2024).

A proposed solution to the high-burden diary surveys is the use of smart surveys. In these surveys, data is collected by sensors available in smartphones. In the case of travel behaviour surveys, the passively collected sensor data can be used to gather information on individual travel behaviour (Yang et al., 2018). The collected geolocation measurements can be used to define segments consisting of trips and stops (Safi et al., 2016), which can then be used to compute statistics on the travel distance, speed, route, and duration. As a result, respondents are no longer required to manually record all the details of their trips, decreasing respondent burden considerably. Furthermore, sensor data may provide more detailed information on travel behaviour that is difficult to obtain from a diary survey, such as the average speed and the exact start and end times of trips (Nahmias-Biran et al., 2018).

In 2018 and 2022/2023, Statistics Netherlands performed large-scale field tests of such an implementation of a smart survey in a travel app (Schouten et al., 2024). A semi-automated approach was used in which the app automatically detects trips and stops, while respondents are asked to annotate these with information on the travel mode and stop purpose (Zahroh et al., expected 2025). Consequently, respondents may still experience the burden of remembering and providing details in the app. One way to further decrease respondent burden would be to automatically detect the transport mode, thereby eliminating the need for respondents to annotate their data. The aim of this research report is to share the preliminary findings of an exploratory study into the feasibility of using a machine learning model to automatically predict the travel mode.

The next section will briefly overview the current state of automatic transport mode detection and the most popular methods used. Section 3 contains information on the data

and pre-processing steps, as well as a description of the methods used to create a preliminary random forest prediction model. After presenting the results, the discussion section will share some considerations and plans for future research.

2 Background

Most automatic transport mode detection methods can be separated into supervised and semi-supervised or unsupervised approaches. Whereas supervised learning algorithms require labeled data to train a prediction model, semi- and unsupervised learning algorithms are often suggested for sparsely labeled or non-labeled data (Sadeghian et al., 2021; Yang et al., 2024). As unlabeled data is typically easier and less expensive to obtain than labeled data, these algorithms are highly valuable in transportation research. However, as they are difficult to evaluate and do not allow for easy comparison (Li et al., 2022; Markos and Yu, 2020), supervised algorithms are often preferred when labeled data is available.

Supervised learning approaches may be further separated into rule-based algorithms, traditional machine learning and deep learning. Rule-based algorithms have been successfully used to predict transport mode on GPS data and have the advantage of high interpretability (Sadeghian et al., 2021; Xiao et al., 2019). However, these approaches often see researchers manually adjusting decision rules, which may be time-intensive and requires expert knowledge (Sadeghian et al., 2021).

Although traditional machine learning models are typically less interpretable, they have been a popular choice for transport mode detection due to their high predictive performance (Yang et al., 2024). Some of the algorithms used in these models are support vector machines (SVM) (Bolbol et al., 2012), naive Bayes (NB) (Nour et al., 2016) and Bayesian networks (BN) (Xiao et al., 2015). Decision tree-based ensemble models like gradient-boosted trees and random forest are also commonly used (Li et al., 2021; Lu et al., 2019; Shafique and Hato, 2016; Liu et al., 2022). Out of these, the random forest model has been particularly popular, as it often outperforms other traditional supervised machine learning algorithms (Bedogni et al., 2016; Bjerre-Nielsen et al., 2020; Hasan et al., 2022; Sadeghian et al., 2022). Considerable advantages of random forests are that they can handle high-dimensional data, require relatively little hyper-parameter tuning to train, and are less prone to over-fitting than many other machine learning methods (Pappu and Pardalos, 2014).

Finally, deep learning algorithms have the advantage of high flexibility in the model input. Whereas traditional supervised learning algorithms often require input on trip segment-level, deep learning algorithms allow for GPS measurement-level input. Creating features on a trip segment-level requires summarising and aggregating information from the GPS measurement-level to trip segment-level, resulting in information loss (Yang et al., 2024). Furthermore, deep learning algorithms such as Recurrent Neural Networks and Long Short-Term Memory have the additional advantage of automatically learning temporal patterns in the data ((Jiang et al., 2023; Yu, 2021; Qin et al., 2019). However, a large disadvantage of these algorithms is that they require a large quantity of high-quality data to be trained successfully, as well as high computational power. This project will therefore focus on traditional machine learning methods.

3 Data and Methods

3.1 Data Collection and Pre-processing

From November 2022 to February 2023, Statistics Netherlands collected smartphone sensor data from a sample of respondents from the general population as part of a field test of a smartphone travel app designed to measure travel behaviour in the Netherlands. The field test had an experimental design, where respondents were asked to use the travel app for one or seven days. Low-quality observations were removed from the data, including geolocation measurements with very low accuracy and respondents for which less than one hour of data was collected. Additionally, all duplicate measurements were removed from the data (Schouten et al., 2024; Gootzen et al., expected 2025).

After removing non-labeled trips, the remaining trip segment data consists of a total of 4219 labeled trips made by 252 users. Each trip segment has a start time, end time, and travel mode label that is one of either walking, bike, e-bike, car, bus, tram, metro, or train. The almost 3 million GPS data points that make up these trip segments consist of a timestamp, longitude, latitude, and accuracy value.

Information on the number of labeled trips for each transport mode class can be found in Table 1. Car trips occur most, followed by walking trips. The public transportation modes metro, tram, and bus occur least frequently in the data.

Table 1: Labeled trips per transport mode class

Car	Bus	E-Bike	Bike	Metro	Tram	Train	Walking
1881	95	146	821	56	34	157	1029

3.2 Feature Engineering

First, features were calculated on the GPS measurement-level. The Haversine formula for great-circle distance between two points was used to determine the distance traveled between two consecutive GPS measurements. The distance traveled and the time difference between points were then used to calculate speed and its higher-order derivatives acceleration, jerk, and snap. Acceleration measures the rate of change of speed, jerk measures the rate of change of acceleration, and snap measures the rate of change of jerk. Additionally, bearing was calculated based on the difference in longitude and latitude between two points. The bearing was only calculated for GPS measurements corresponding with a speed of at least 0.6 m/s. Calculating the bearing for GPS measurements with a lower speed resulted in highly diverging consecutive bearing values that were deemed uninformative. All calculations were performed with the `geosphere` package in R (Hijmans, 2024; R Core Team, 2024).

In the second step, GPS-level features were aggregated to the trip segment-level by summarising the GPS-based features for all data points between the trip segment's start and end time. Additionally, the accuracy of GPS measurements, frequency of measurement gaps (more than 5 minutes between consecutive timestamps), and frequency and size of outliers and implausible values were taken into account. Lastly, features such as day of the week, time of day, and whether the trip was made during rush hour were added to the model.

3.3 Modeling

The preliminary model is a random forest classifier based on the features calculated from the GPS data described previously. As preparation, the data was split into training data and testing data using a 70/30 split ratio. A blocking strategy was used to ensure that all trip segments belonging to a single respondent are either included in the training or testing data. This strategy should prevent overly optimistic predictions when potentially similar trips from a single respondent are used for both training and testing the model.

A random forest prediction model was created with the `ranger` package in R, using

the streamlining functions included in the `caret` package (Wright and Ziegler, 2017; Kuhn and Max, 2008). Model hyper-parameters were tuned using 10-fold cross-validation on the training data. The same type of blocking strategy that was used for splitting the train and test data was used for the creating the folds. The number of decision trees in the random forest model was set to 50 trees and 54 features were considered by each tree when splitting a node.

3.4 Evaluation

The model performance is measured with the balanced accuracy and F1 score of the test data for each of the transport modes. Both evaluation criteria are confusion matrix-based metrics that are suitable for multi-class classification in imbalanced datasets (Luque et al., 2019). Whereas the balanced accuracy is the mean of the true positive rate (recall) and the true negative rate, the F1 score balances the precision or positive predictive value with the recall. F1 scores close to 1 indicate high precision and recall, whereas scores below 0.5 indicate a poor prediction performance.

The F1 Score is the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The balanced accuracy is the average of recall for each class:

$$\text{Balanced Accuracy} = \frac{1}{C} \sum_{i=1}^{C} \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}$$

where C is the number of classes.

Additionally, feature importance in the random forest model is evaluated using the permutation variable importance measure (Altmann et al., 2010). This measure can be understood as the mean decrease in prediction accuracy and can be calculated by looking at the decrease in accuracy when the values on one feature are randomly shuffled, thereby isolating the effect of that particular feature on the model's prediction performance. This method is often preferred over the default impurity or Gini importance measure (Janitza et al., 2018).

Table 2: Confusion matrix comparing observed and predicted transport modes in the test data

	Observed							
Predicted	Car	Bus	E-bike	Bike	Metro	Tram	Train	Walking
Car	**472**	33	4	30	11	8	8	19
Bus	1	**0**	0	3	1	1	0	0
E-bike	1	0	**3**	10	0	0	0	0
Bike	4	0	11	**170**	0	1	0	19
Metro	0	0	0	0	**1**	0	0	0
Tram	1	0	0	0	0	**0**	0	0
Train	9	0	0	1	3	0	**31**	0
Walking	17	0	0	13	8	0	0	**279**

Table 3: F1 scores and accuracy per class in the test data. High values indicate good predictive performance

	Car	Bus	E-bike	Bike	Metro	Tram	Train	Walking
F1 score	0.87	0	0.19	0.79	0.08	0	0.75	0.88
Accuracy	0.88	0.50	0.58	0.86	0.52	0.50	0.89	0.92

4 Preliminary Results

4.1 Prediction Performance

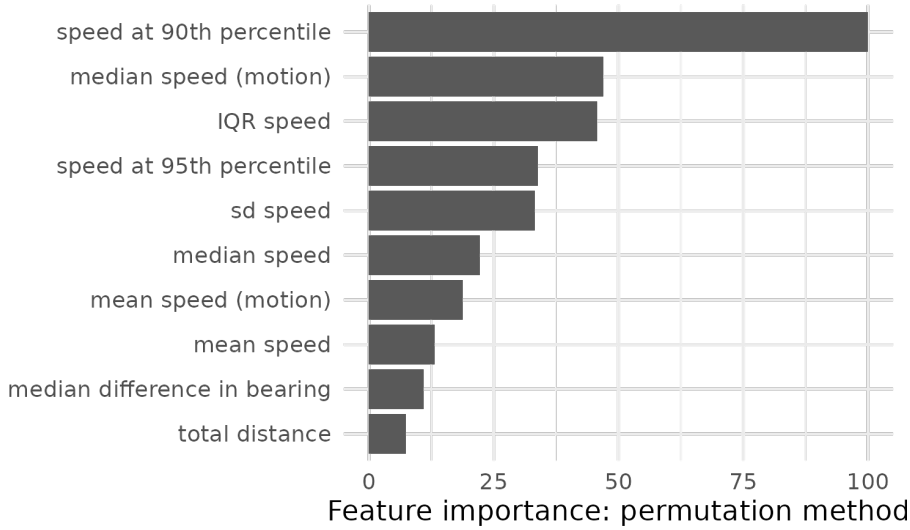
The model's prediction performance on the test data is summarised in the confusion matrix and evaluation criteria overview in Tables 2 and 3. Both tables highlight considerable differences in prediction performance across transport modes. The overrepresented classes walking, car, and bike are often classified correctly, achieving F1 scores of 0.88, 0.87, and 0.79 respectively. The model performed moderately well on train trips, with some misclassification into car trips and an F1 score of 0.75.

Contrastingly, the model did not perform well on the underrepresented classes bus, metro, tram, and e-bike. Bus and tram trips were consistently misclassified as car trips, while metro trips were also misclassified as walking trips. Furthermore, over 60% of e-bike trips were misclassified as bike trips. F1 scores for these transport modes are all below 0.2. Bus and tram received F1 scores of 0, as these were never correctly classified.

4.2 Feature Importance

Feature importance analysis revealed that speed-related features were the most important predictors in the random forest model, as they collectively represent the top 8 most influential

Figure 1: Feature importance of the 10 most important features



features. The top 10 most important features out of 114 features included in the model are provided in Figure 1.

The important speed-related features describe the central tendency and variability of a trip's speed, such as the overall mean/median and the mean/median speed when a person is in motion, the standard deviation and the inter-quartile range. The most influential feature is speed at the 90th percentile, which represents the value below which 90% of the GPS data points in a trip fall. It can be interpreted as the speed value that separates the highest 10% of speed observations during a trip.

The importance of speed-related features aligns with the model's performance in distinguishing walking, car, and bike trips, which tend to have distinct speed patterns. However, it may also partially explain the poor prediction performance for e-bike, tram, and bus. These transport modes may have speed characteristics that are difficult to distinguish from bike and car trips. As they are also underrepresented in the data, it may be difficult for a machine learning model to learn the unique speed patterns associated with these transport modes.

5 Discussion

In this preliminary random forest model, sensor data collected in a travel app were used to create features to predict the transport mode of trips made by respondents. The reported

findings are meant to serve as a baseline for further research to see whether prediction performance can be increased by adding additional data sources and accounting for temporal dependencies. While the preliminary random forest model is able to predict common transport modes such as car, bike, and walking relatively well, it struggles to classify the underrepresented transport modes such as bus, tram, and metro. This discussion will outline some considerations and steps to be taken in the remainder of this project.

5.1 Additional Data Sources

First, additional features will be created that incorporate geospatial information using OpenStreetMap (OSM) data. Researchers have previously incorporated geospatial information into their transport mode detection models by creating features such as the proximity to public transport stations and public transport network lines (Liu et al., 2022; Li et al., 2021; Nour et al., 2016). Adding features based on geospatial information from an external source like OSM is expected to help in distinguishing between public transport modes and non-public transport modes that share similar speed characteristics.

5.2 Accounting for Temporal Dependencies

Random forest models assume that observations are independent and do not automatically account for potential temporal dependencies between observations (Hu and Szymczak, 2023; Sela and Simonoff, 2020). As the data in this project contains multiple trips made by individuals over a time period, it may be worth exploring whether we can account for potential temporal dependencies between trips by adding lagged features. Lagged features contain information on previous trips made at earlier time points. Information may include the (predicted) transport mode of previous trips, but also the time between consecutive trips and whether previous trips were made during a similar time of day. For the current test data, the information on previous trips is available and can be used directly as input for the features. However, as this information will not be available for unlabeled future data, transport mode prediction will become an iterative process with earlier predictions serving as input to help predict future trips.

5.3 Generalisability

When a machine learning model is trained on data situated in a specific time and context and used to make predictions on unseen future data, there may be concerns about the generalisability of the trained model. This concern may be relevant for the preliminary model that has been created, as well as for future models that are trained on the same data. As the data was collected exclusively during the winter of 22/23, we do not know whether the model may be biased toward seasonal travel patterns caused by differences in weather or events. Travel behaviour of individuals may differ considerably depending on the season. This may especially be an issue for the transport modes that are underrepresented in the data, as the risk of issues with extrapolation tends to be larger for these classes. Potentially testing the model on travel behaviour data collected in a different time period may provide some insights into whether generalisability will be a large concern.

References

Altmann, A., Tološi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347.

Bedogni, L., Di Felice, M., & Bononi, L. (2016). Context-aware android applications through transportation mode detection techniques. *Wireless communications and mobile computing*, *16*(16), 2523–2541. https://doi.org//10.1002/wcm.2702

Bjerre-Nielsen, A., Minor, K., Sapieżyński, P., Lehmann, S., & Lassen, D. D. (2020). Inferring transportation mode from smartphone sensors: Evaluating the potential of Wi-Fi and Bluetooth. *PLOS ONE*, *15*(7), e0234003. https://doi.org/10.1371/journal.pone.0234003

Bolbol, A., Cheng, T., Tsapakis, I., & Haworth, J. (2012). Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. *Computers, Environment and Urban Systems*, *36*(6), 526–537. https://doi.org/10.1016/j.compenvurbsys.2012.06.001

Gootzen, Y., Klingwort, J., & Schouten, B. (expected 2025). *Data quality aspects for location tracking in smart travel and mobility surveys* [Discussion Paper].
Statistics Netherlands.

Hasan, R. A., Irshaid, H., Alhomaidat, F., Lee, S., & Oh, J.-S. (2022). Transportation Mode Detection by Using Smartphones and Smartwatches with Machine Learning. *KSCE Journal of Civil Engineering*, *26*(8), 3578–3589. https://doi.org/10.1007/s12205-022-1281-0

Hijmans, R. J. (2024). *Geosphere: Spherical trigonometry* [R package version 1.5-20]. https://github.com/rspatial/geosphere

Hu, J., & Szymczak, S. (2023). A review on longitudinal data analysis with random forest. *Briefings in Bioinformatics*, *24*(2), bbad002. https://doi.org/10.1093/bib/bbad002

Janitza, S., Celik, E., & Boulesteix, A.-L. (2018). A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, *12*(4), 885–915.

Jiang, Z., Huang, A., Qi, G., & Guan, W. (2023). A Framework of Travel Mode Identification Fusing Deep Learning and Map-Matching Algorithm. *IEEE Transactions on Intelligent Transportation Systems*, *24*(6), 6401–6415. https://doi.org/10.1109/TITS.2023.3250660

Kuhn & Max. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

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. https://doi.org/10.26599/TST.2020.9010014

Li, Z., Xiong, G., Wei, Z., Lv, Y., Anwar, N., & Wang, F.-Y. (2022). A Semisupervised End-to-End Framework for Transportation Mode Detection by Using GPS-Enabled Sensing Devices. *IEEE Internet of Things Journal*, *9*(10), 7842–7852. https://doi.org/10.1109/JIOT.2021.3115239

Liu, Y., Miller, E., & Habib, K. N. (2022). Detecting transportation modes using smartphone data and GIS information: Evaluating alternative algorithms for an integrated smartphone-based travel diary imputation. *Transportation Letters*, *14*(9), 933–943. https://doi.org/10.1080/19427867.2021.1958591

Lu, Z., Long, Z., Xia, J., & An, C. (2019). A Random Forest Model for Travel Mode Identification Based on Mobile Phone Signaling Data. *Sustainability*, *11*(21), 5950. https://doi.org/10.3390/su11215950

Luque, A., Carrasco, A., Martín, A., & de Las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, *91*, 216–231. https://doi.org/10.1016/j.patcog.2019.02.023

Markos, C., & Yu, J. J. (2020). Unsupervised deep learning for gps-based transportation mode identification. *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 1–6. https://doi.org/10.1109/ITSC45102.2020.9294673

McCool, D., Lugtig, P., Mussmann, O., & Schouten, B. (2021). An App-Assisted Travel Survey in Official Statistics: Possibilities and Challenges. *Journal of Official Statistics*, *37*(1), 149–170. https://doi.org/10.2478/jos-2021-0007

Nahmias-Biran, B.-h., Han, Y., Bekhor, S., Zhao, F., Zegras, C., & Ben-Akiva, M. (2018). Enriching Activity-Based Models using Smartphone-Based Travel Surveys. *Transportation Research Record*, *2672*(42), 280–291. https://doi.org/10.1177/0361198118798475

Nour, A., Hellinga, B., & Casello, J. (2016). Classification of automobile and transit trips from Smartphone data: Enhancing accuracy using spatial statistics and GIS. *Journal of Transport Geography*, *51*, 36–44. https://doi.org/10.1016/j.jtrangeo.2015.11.005

Pappu, V., & Pardalos, P. M. (2014). High-dimensional data classification. *Clusters, Orders, and Trees: Methods and Applications: In Honor of Boris Mirkin's 70th Birthday*, 119–150.

Prelipcean, A. C., Susilo, Y. O., & Gidófalvi, G. (2018). Collecting travel diaries: Current state of the art, best practices, and future research directions. *Transportation Research Procedia*, *32*, 155–166. https://doi.org/10.1016/j.trpro.2018.10.029

Qin, Y., Luo, H., Zhao, F., Wang, C., Wang, J., & Zhang, Y. (2019). Toward Transportation Mode Recognition Using Deep Convolutional and Long Short-Term Memory Recurrent Neural Networks. *IEEE Access*, *7*, 142353–142367. https://doi.org/10.1109/ACCESS.2019.2944686

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Sadeghian, P., Håkansson, J., & Zhao, X. (2021). Review and evaluation of methods in transport mode detection based on GPS tracking data. *Journal of Traffic and Transportation Engineering (English Edition)*, *8*(4), 467–482. https://doi.org/10.1016/j.jtte.2021.04.004

Sadeghian, P., Zhao, X., Golshan, A., & Håkansson, J. (2022). A stepwise methodology for transport mode detection in GPS tracking data. *Travel Behaviour and Society*, *26*, 159–167. https://doi.org/10.1016/j.tbs.2021.10.004

Safi, H., Assemi, B., Mesbah, M., & Ferreira, L. (2016). Trip Detection with Smartphone-Assisted Collection of Travel Data. *Transportation Research Record*, *2594*(1), 18–26. https://doi.org/10.3141/2594-03

Schouten, B., Remmerswaal, D., Elevelt, A., Groot, J., Klingwort, J., Schijvenaars, T., Schulte, M., & Vollebregt, M. (2024). *A smart travel survey: Results of a push-to-smart field experiment in the netherlands*. https://doi.org/10.13140/RG.2.2.30248.38404

Sela, R. J., & Simonoff, J. S. (2020). RE-EM trees: A data mining approach for longitudinal and clustered data. *Mach Learn*, *86*, 169–207. https://doi.org/DOI10.1007/s10994-011-5258-3

Shafique, M. A., & Hato, E. (2016). Travel Mode Detection with Varying Smartphone Data Collection Frequencies. *Sensors*, *16*(5), 716. https://doi.org/10.3390/s16050716

Wang, K., Liu, Y., Hossain, S., & Nurul Habib, K. (2024). Who drops off web-based travel surveys? Investigating the impact of respondents dropping out of travel diaries during online travel surveys. *Transportation*, 1–37. https://doi.org/10.1007/s11116-024-10510-8

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Xiao, G., Cheng, Q., & Zhang, C. (2019). Detecting travel modes using rule-based classification system and gaussian process classifier. *IEEE Access*, *7*, 116741–116752.

Xiao, G., Juan, Z., & Zhang, C. (2015). Travel mode detection based on GPS track data and Bayesian networks. *Computers, Environment and Urban Systems*, *54*, 14–22. https://doi.org/10.1016/j.compenvurbsys.2015.05.005

Yang, N., Al Haddad, C., Yamnenko, I., & Antoniou, C. (2024). Machine Learning for Data-Centric Transport Mode Detection: A Systematic Review. https://doi.org/10.2139/ssrn.4960556

Yang, X., Stewart, K., Tang, L., Xie, Z., & Li, Q. (2018). A review of gps trajectories classification based on transportation mode. *Sensors*, *18*(11), 3741.

Yu, J. J. Q. (2021). Travel Mode Identification With GPS Trajectories Using Wavelet Transform and Deep Learning. *IEEE Transactions on Intelligent Transportation Systems*, *22*(2), 1093–1103. https://doi.org/10.1109/TITS.2019.2962741

Zahroh, S., Lugtig, P., Gootzen, Y., Klingwort, J., & Schouten, B. (expected 2025). *Predicting trip purpose in a smartphone-based travel survey* [Discussion Paper]. Statistics Netherlands.