# Abstract
# EMOS Master thesis competition 2025

# ‘Predicting Travel Purpose in a Smartphone-Based Travel Survey ’

Author: **Solichatus Zahroh, Utrecht University**

Keywords: GPS, smartphone-based travel survey, travel purpose prediction

## 1. INTRODUCTION

A travel survey records people's movement patterns in a specific area and was valuable in many research fields. By analysing some supplementary variables, one could uncover distinctive patterns demonstrated by travellers, regardless of the distance or the duration of their journey. However, the general population travel survey was burdensome for the respondents as each respondent should document a full-day trip, including the precise start times, end times, and locations. Predicting travel purposes automatically would be beneficial since it is difficult to accurately recall, and report stops and tracks from memory in traditional travel surveys.

In travel behaviour analysis, trip purpose was a fundamental yet complex research issue. Understanding the meanings of activities within the context of a trip was often essential. Most existing methods, however, depend on obtaining sensitive information from passengers, such as their home addresses or daily travel logs from surveys, to generate precise conclusions. Consequently, it was seldom applicable in real-world circumstances due to the reluctance of certain respondents to offer the data (Liao et al., 2022).

The growing popularity of smartphones was enabling the emergence of smart city applications (Soares et al., 2019). Smartphones used various sensors, including the Global Positional System (GPS), Global System for Mobile Communications (GSM), and accelerometer, to gather real-time data for location tracking and detection (Yu et al., 2012). The collection of GPS sensor data was advantageous for discovering movement patterns and portraying movement behaviour (Calabrese et al., 2013).

Statistics Netherlands (Centraal Bureau voor de Statistiek or CBS) implemented an innovation in travel survey data collection in 2020 (McCool et al., 2021). Automatically collecting GPS sensor data could alleviate the burden on respondents by eliminating the necessity for them to remember and report all of their daily trips. Using smart devices to track visit times and stop durations also enabled the analysis of passengers' behaviours, which is critical for predicting the purpose of their trip (Kakar, 2020).

An automated travel diary was created for the survey respondents based on the GPS sensor data. Each day was divided into distinct segments representing stationary periods (stops) and periods of travel (tracks). Through the personal movement investigation data gathered by multi-day use of the CBS app, movement patterns of travellers could be uncovered. Such an app-based approach could reduce recall bias, collected a lot of data with minimal effort and was less time-consuming (Zhou et al., 2022). The travel survey community, therefore, considered that GPS data would emerge as a key method for future data collection, offering a solution to current challenges (Bricka et al., 2012).

The current approach was to simply ask someone why they travelled. Detailed activity-related data, such as nearby places and past choices made by other travellers, was absent from the current techniques (Cui et al., 2018). Examining the historical places people have visited could replicate the personalised preferences of travel destinations. Furthermore, publicly accessible information about nearby places, such as Open Street Map (OSM) data, could provide a more comprehensive insight into the functionalities of a certain site.

Online resources like OSM and Google Places API, as well as offline land use data collected by CBS, were the two primary sources of information regarding POIs. This dataset contained information regarding specific geographical places within a given area. Google Places API requests deliver nearby search, text search, radar search, and place details queries. In contrast, OSM provides free access with limited coverage and a more lenient license compared to Google Places. OSM relies on crowd-sourced data and has strong community support. Additionally, OSM may be downloaded and used offline. However, there was no standardized format of using OSM therefore making it impossible to verify the accuracy of OSM.

## 2. OBJECTIVE

The goal of this study is, therefore, to use GPS data, external spatial and temporal patterns data, and socio-demographic characteristics to automate trip purpose prediction. It is expected that it is no longer necessary to directly ask respondents about the purpose of their trip. One could gain a more profound understanding of travel patterns within a particular geographic area and time frame.

This study aims to answer the general research question “How well can we predict the travel purpose using sensor data from a smartphone-based travel diary study?” which can be translated into two sub-questions as follows:

1. To what extent are external spatial and temporal patterns data helpful in predicting travel purposes?

2. To what extent do individual behaviours and characteristics influence the accuracy of travel purpose prediction?

## 3. METHODS

To answer the aim of this study, several machine learning models were used. Before the training process, several data cleaning was being done. The collected data consisted of 21,397,699 location observations. Multiple observations were obtained at one location within a time frame of approximately 5 seconds, and the most accurate set of location measurements was selected. After eliminating redundant and low-quality data, there were 12677 locations from 456 users. About half

of the observations were classified as stops (stationary period) and the other half were tracks (moving period).

The information about the distance of each track and the duration of the stops and tracks were added as features, along with the time of day (in 24 hours) and day of the week of each stop and track. Since the total number of visits to regularly visited places, such as one's house, was rather low, numerous stop locations in close proximity to one another were merged, resulting in an accuracy of 55 meters.

Multiple bounding boxes with varying radiuses were determined from OSM for different tags associated with trip purposes. Four distinct radiuses—25, 35, 50, and 200 metres—were given to evaluate discrepancies that may arise when a stop was located at a position that did not align with OSM data. A large number of features could not be solely dropped into the model since the information of four distinct radiuses of the bounding box contains overlapping information. In order to maximise training time, reduce noisy attributes, and prevent overfitting the data, the model was constrained to only one radius per tag.

The analysis excluded data with insufficient quality, which included less than one hour and less than 2000 observations. Furthermore, noisy data that might adversely affect the algorithm's stop-track categorization was excluded using median smoothing by manual and automatic detection. Data acquired beyond the specified reference periods, which might be either one or seven days, and tracks more than 24 hours were also omitted. The locations outside the Netherlands were also excluded. The final dataset for training process consisted of 4961 locations (Figure 1).

This dataset was divided into training and testing sets with a ratio of 80:20. Due to the complexity of the data, neural network model as one of the machine learning models was used to analyse the data. However, given the diverse sources of error and uncertainty involved in the study, it may not be the most effective technique to choose a single model for both development and application (Cheng et al., 2019).

For comparison, various machine learning models, including Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGB), were evaluated. The test sets of various models were assessed using two metrics, specifically balanced accuracy and F1-score.

## 4. RESULTS

The majority of stops occurred between the hours of 8 a.m. and 4 p.m. coinciding with typical working hours. Home accounted for most stops (30.3%), while education (0.7%), sport (1.2%), and transit (1.4%) had the fewest stops. The car was the most prevalent form of transportation, accounting for 33.7% of the tracks, while the tram was the least preferred option, representing only 0.6%.

The initial analysis was performed without OSM data, and only included GPS data and socio-demographic variables (gender, age, income group, household numbers, household types, socio-economic groups, car ownership, and working hours). The training and testing data had approximately 65% accuracy. The balanced accuracy of the assessment results ranged from 53.9% for the education category to 90.9% for the home category. The test's accuracy decreased by 12% when it incorporated data from OSM within four distinct radiuses. In the subsequent trial, the tags were categorized and examined as count data and percentages. A percentage represents the ratio of the total number of POIs in a specific category to the total number of all POIs. This was done since

tag selection was previously based on the labels of trip purposes. Various combinations of category-only, percentage-only, category and individual tags, and percentage and individual tags were utilized. Only the model accuracy of the percentage-only model decreased compared to the prior model.

The optimal model was an ANN model that utilized OSM data, using both percentage and individual tags. In addition, weather information was incorporated to enhance the model. Regrettably, the inclusion of weather data did not enhance the performance of the model. The RF, XGB, SVM, and NB models were trained using the same set of training control parameters and data sets for comparison. The RF model's accuracy was perfect for the training model, but not for the testing model. The XGB model also achieved near-perfect training accuracy, given that both models are tree-based. XGB's accuracy was greater than 90% for the training set, but not for the test set. The SVM model produced similar balanced accuracy compared to the ANN. Considering that SVM was initially developed to enhance the training of ANN. Unfortunately, the NB model performed poorly. The classifier exhibited a lack of ability to differentiate between classes and produced almost random predictions, leading to an accuracy rate of approximately 50% (Table 1). Table 2 shows the confusion matrix of ANN and XGB models as the best model (similar balanced accuracy for the test data). From confusion matrix, we can see that “other” category mostly misclassified as visit or shop and “pick-up” had lower accuracy despite of low number of misclassification due to small sample size.

In conclusion, a smartphone-based travel diary study could predict the travel purpose pretty well. The model’s accuracy was a bit lower than in past studies due to its ability to classify a greater number of trip-purpose labels than in previous research. The model achieved satisfactory accuracy on the initial attempt when trained without OSM data, and some sociodemographic factors were essential. However, only the respondents' ages, which serve as indicators of individual qualities, became significant factors in the integration of OSM data. The overall visit frequency to the same area was crucial, representing 45% in the ANN model and 25% in the XGB model. This demonstrates that a data collection period of seven days was better than one day. It demonstrated individuals' ability to adjust and thrive in various situations. The study concludes that individual behaviors (visiting the same location with the same purpose) were more accurate predictors of travel purpose than individual characteristics (administrative data).

Stop duration was the sole factor responsible for all variability in all models. The same location might serve multiple purposes, depending on the length of the visit. This suggested that we can use temporal patterns to identify the purpose of the trip. In OSM, the number of recreational facilities were important data. The presence of a higher number of shops and sports facilities within a 25-metre radius significantly enhanced the probability of making a halt. Spatial data aided in predicting the purpose of trips.

Recording spatial and temporal patterns diminished the importance of certain characteristics. This was a promising indication for CBS to accurately predict a trip's purpose in real-time. The absence of a request for the users' sociodemographic characteristics prevented immediate access to this information in real-time. Nevertheless, people might visit the same location several times on different occasions. In order to accurately determine the purpose of the trip, it was important to include additional information, such as the respondents' occupation or their participation in hobbies-related memberships. Utilizing spatial and temporal patterns was valuable for predicting trip purposes, and individual behaviors had minimal influence on the accuracy of trip-purpose prediction. Some types of stops would be more important for making accurate predictions than others.

Based on the challenges we encountered, this study offered several recommendations for future research. Firstly, the data collection time should not be limited to one season, data collection for the whole year can be an option. Secondly, due to timing errors, inaccuracies in both the satellite and receiver clocks, as well as relativity effects, can result in position errors of up to two meters, more than one observation should be selected per location. Thirdly, indoor and outdoor activities should be divided and more sample sizes per class are needed to reach model convergence. Additionally, certain information, such as respondents' profession and hobbies-related subscriptions, should be included to predict specific purposes. Lastly, exploration of individual characteristics, such as the tendency for teachers to visit educational places more frequently than others, must be done. Moreover, we might improve the model by opting for better tags in OSM and leveraging the most up-to-date OSM data. The predicted increase in the quantity of data points was expected to facilitate the training of more complex models, hence enabling the detection of variations in unique behaviors across many seasons and the identification of weather dependencies.

## 4. CONTRIBUTION

Sensor data from a smartphone-app travel diary app successfully predicted the purpose of a trip by utilising spatial and temporal patterns. Prediction models gave more importance to trends that occur over time in specific locations. The best model from this study can be used a prediction model for the future CBS travel survey app. It is not necessary to ask the private information from respondents since the temporal and spatial features are useful to predict the travel purposes automatically. Some types of stops would be more important for making accurate predictions than others, for instance predicting shop is more interesting than home because home can easily be predicted due to high number of daily visits. The chosen tags from OSM in this study were trained from several trials therefore it can be considered as good tags to predict travel purposes. Adding land-use information from CBS database might also be useful in predicting travel purposes as additional information besides OSM.

## TABLES AND FIGURES

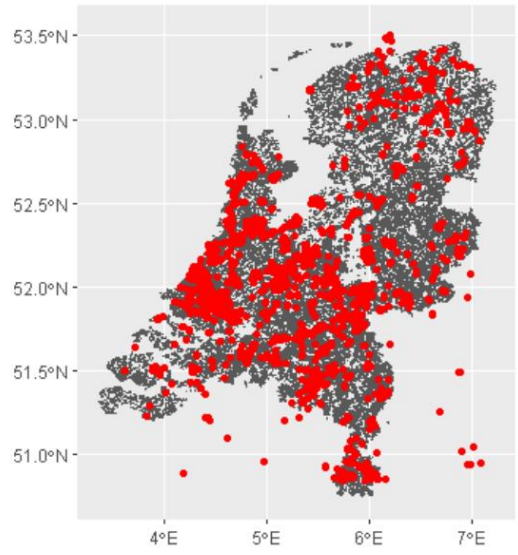


**Figure 1.** Mapping of the Location Data

**Table 1.** Balanced Accuracy of Test Model (in %)

| Model | ANN | RF | XGB | SVM | NB |
|---|---|---|---|---|---|
| **Overall** | 72.3 | 77.5 | 77.7 | 70.8 | 42.7 |
| **Pick-up** | 75 | 78.2 | 75.3 | 70.7 | 50 |
| **Edu** | 83.3 | 80.3 | 83.1 | 80.4 | 50 |
| **Others** | 62.9 | 69 | 74.1 | 64.4 | 50 |
| **Transit** | 80.1 | 79.1 | 83.3 | 71.7 | 50 |
| **Sport** | 71.1 | 73.1 | 81.2 | 68.1 | 50 |
| **Home** | 91.3 | 92.8 | 93.2 | 90.7 | 52.6 |
| **Visit** | 67.5 | 62.3 | 74.6 | 67.3 | 50 |
| **Work** | 84.1 | 90.8 | 87.9 | 83.9 | 49.9 |
| **Shop** | 81.5 | 86 | 83.8 | 80.1 | 55.5 |

ANN model is the ANN_6 model (ANN model with OSM data from 1 radius as count data and percentage without weather data). The other models represent the best models with the optimum parameters tuning.

**Table 2.** Confusion Matrix of the Best Model

**ANN Model**

| Obs / Pred | Pick-up | Edu | Other | Transit | Sport | Home | Visit | Work | Shop |
|---|---|---|---|---|---|---|---|---|---|
| **Pick-up** | **41** | 0 | 4 | 0 | 1 | 5 | 4 | 3 | 7 |
| **Edu** | 0 | **12** | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **Others** | 9 | 1 | **22** | 3 | 3 | 3 | 7 | 4 | 10 |
| **Transit** | 2 | 0 | 1 | **21** | 0 | 0 | 1 | 8 | 3 |
| **Sport** | 3 | 0 | 3 | 0 | **13** | 0 | 1 | 4 | 0 |
| **Home** | 4 | 2 | 8 | 2 | 2 | **377** | 15 | 14 | 12 |
| **Visit** | 3 | 1 | 10 | 0 | 3 | 11 | **24** | 2 | 4 |
| **Work** | 10 | 1 | 6 | 5 | 4 | 7 | 3 | **116** | 4 |
| **Shop** | 6 | 1 | 19 | 3 | 4 | 3 | 7 | 7 | **88** |

**XGB Model**

| Obs / Pred | Pick-up | Edu | Other | Transit | Sport | Home | Visit | Work | Shop |
|---|---|---|---|---|---|---|---|---|---|
| **Pick-up** | **42** | 0 | 4 | 1 | 1 | 5 | 4 | 6 | 8 |
| **Edu** | 0 | **12** | 0 | 0 | 2 | 1 | 0 | 2 | 0 |
| **Others** | 8 | 1 | **38** | 1 | 3 | 4 | 5 | 3 | 11 |
| **Transit** | 1 | 0 | 2 | **23** | 0 | 0 | 0 | 1 | 5 |
| **Sport** | 6 | 0 | 2 | 0 | **19** | 0 | 0 | 0 | 0 |
| **Home** | 3 | 1 | 3 | 2 | 2 | **384** | 17 | 13 | 6 |
| **Visit** | 3 | 1 | 7 | 0 | 1 | 5 | **32** | 2 | 4 |
| **Work** | 6 | 2 | 7 | 5 | 1 | 5 | 4 | **127** | 3 |
| **Shop** | 9 | 1 | 10 | 2 | 1 | 2 | 0 | 5 | **91** |

## 6. REFERENCES


[1] Atwal, K. S., Anderson, T., Pfoser, D., & Züfle, A. (2022). Predicting building types using OpenStreetMap. Scientific Reports, 12(1), 19976. https://doi.org/10.1038/s41598-022-24263-w

[2] Bricka, S. G., Sen, S., Paleti, R., & Bhat, C. R. (2012). An analysis of the factors influencing differences in survey-reported and GPS-recorded trips. Transportation Research Part C: Emerging Technologies, 21(1), 67–88. https://doi.org/10.1016/j.trc.2011.09.005

[3] Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation Research Part C: Emerging Technologies, 26, 301–313. https://doi.org/10.1016/j.trc.2012.09.009

[4] 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. https://doi.org/10.1016/j.tbs.2018.09.002

[5] Cui, Y., Meng, C., He, Q., & Gao, J. (2018). Forecasting current and next trip purpose with social media data and Google Places. Transportation Research Part C: Emerging Technologies, 97, 159–174. https://doi.org/10.1016/j.trc.2018.10.017

[6] Kakar, A. (2020). Trip Purpose and Prediction. International Journal of Engineering Research, 9(10), 278–285.

[7] Liao, C., Chen, C., Guo, S., Wang, Z., Liu, Y., Xu, K., & Zhang, D. (2022). Wheels Know Why You Travel: Predicting Trip Purpose via a Dual-Attention Graph Embedding Network. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 6(1), 1–22. https://doi.org/10.1145/3517239

[8] 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

[9] Soares, E., Revoredo, K., Baiao, F., A. De M. S. Quintella, C., & V. Campos, C. A. (2019). A Combined Solution for Real-Time Travel Mode Detection and Trip Purpose Prediction. IEEE Transactions on Intelligent Transportation Systems, 20(12), 4655–4664. https://doi.org/10.1109/TITS.2019.2905601

[10] Zhou, Y., Zhang, Y., Yuan, Q., Yang, C., Guo, T., & Wang, Y. (2022). The Smartphone-Based Person Travel Survey System: Data Collection, Trip Extraction, and Travel Mode Detection. IEEE Transactions on Intelligent Transportation Systems, 23(12), 23399–23407. https://doi.org/10.1109/TITS.2022.3207198