Discussion Paper

Predicting trip purpose in a smartphone-based travel survey

Solichatus Zahroh
Peter Lugtig
Yvonne Gootzen
Jonas Klingwort
Barry Schouten

May 2025

Abstract

Population-wide travel surveys are conducted to investigate individuals' patterns of traveling. These surveys are often burdensome. The widespread use of smartphones allows for the collection of trip data without relying on traditional travel survey diary responses. Location monitoring data can be utilized to split the day into segments where someone is traveling (track), or is stationary (stop). Currently, respondents have to label each trip and stop, what mode of transport they travel with, and what the purpose of a stop is (e.g. shopping, work, school, home). By integrating GPS data from smartphones with administrative data and temporal and spatial data, this paper studies to what degree it is possible to automatically predict the trip purpose. Multiple machine-learning models were trained and evaluated to unveil the effectiveness of stop-purpose prediction. In late 2022, Statistics Netherlands collected GPS data that contained 12 distinct labels denoting the purposes of the trips. The most optimal artificial neural network model and extreme gradient boosting technique obtained a balanced accuracy of 90% for the purpose of being at home. Primarily, classes that included only a small number of observations were erroneously categorized as classes with a large number of observations. Administrative data do not help to improve model prediction beyond spatiotemporal covariates. Increasing the duration of data gathering substantially enhanced the precision of the model. To summarize, smartphone-based travel data has considerable potential as a data source for trip purpose prediction but cannot yet be used to predict trip purpose automatically.

Keywords: GPS, machine learning, neural network, location tracking, geo-data, sensor data, digital behavioral data, diary surveys, mobility patterns, response burden

Contents

1 Introduction

Time-use surveys aim to record people's lifestyles and have many applications. Respondents keep a diary for one or more days up to a week. Time-use surveys are strong candidates to go smart. One reason is that they are burdensome; events need to be reported at a relatively high time frequency. Another reason is that they require recall of starting and end times, which is particularly problematic for shorter-duration activities. A last reason is that definitions of what activities are and how to categorize them may be difficult for the average respondent.

Prominent examples are travel or passenger mobility surveys. The focus shifts to movement patterns in a specific area. Typically, travel surveys employ large general population samples to inform policy-makers about the diversity in travel patterns across regions. They often ask for very detailed information on the nature and purposes of travel. Precise start times, end times, and locations need to be reported and supplemented with modes of travel and purposes. Consequently, within the realm of time-use surveys, travel surveys probably benefit the most from added smart features. Predicting travel modes and trip purposes semi-automatically would be two such features. In this paper, we focus on the purpose of the trip.

Complete data on a diverse set of respondents is necessary to analyze people's mobility's spatial and temporal patterns and understand their travel behavior (Omrani 2015). Several interconnected variables, including an increase in household incomes and car ownership, as well as the emergence of part-time and remote work opportunities that offer flexible schedules and locations, caused spatiotemporal variations in recent decades (Bohte et al. 2008). So, there is both an interest in more detailed travel pattern data by authorities managing infrastructure as a limit to what can be asked.

Smart devices may offer a solution. Smartphones' growing popularity enables the emergence of smart city applications (Soares et al. 2019). Smartphones use 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 (Xiao et al. 2012). Collecting GPS sensor data is advantageous for discovering movement patterns and portraying movement behavior (Calabrese et al. 2013).

Recent smartphone-based travel studies have allowed the recording of locations and time more precisely than traditional diary surveys, which often rely on zip codes for specifying locations. 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 enables the analysis of passengers' behaviors, which is required for predicting the purpose of their trip (Kakar 2020).

It remains, however, a challenge to determine why a person is at a specific time in a particular place (Stopher et al. 2005). Research and literature on trip purpose prediction still is relatively rare compared to travel mode prediction (Nguyen et al. 2020; Omrani 2015; Shafique and Hato 2016; Soares et al. 2019; Zhao et al. 2020). To date, respondents are usually asked about the purpose of the trip, both in traditional diaries and in smartphone-based studies.

Often, trip purpose prediction methods rely on sensitive information supplemented by respondents, such as their home and work addresses, to generate precise conclusions. In practice, respondents are reluctant to offer the data (Liao et al. 2022).

Trip purpose can be predicted using different types of features. A natural set of features that can be derived directly from location data segmentation are the stop's duration and start time and preceding trips leading to the stop. Other features, such as land use and socio-demographic background of respondents, have also been explored as potential explanatory variables. (Cui et al. 2018). Trip purpose classifications are hierarchical, and the depth of the hierarchy considered in training is influential, as may be expected (Jin et al. 2022). Xiao et al. (2016) achieved an accuracy rate of 95% in the classification of six different trip purposes using artificial neural networks by utilizing many features, such as point-of-interest (POI) data, polygon-based information, socio-demographic background, and travel mode. Oliveira et al. (2014) applied a decision tree model to predict 12 different trip purposes and obtained 65% accuracy by employing land-use, socio-demographic, and trip-related data as features. In their study, Feng and Timmermans (2015) used 11 trip purpose categories using POIs, start time, duration, and transport modes as features and achieved accuracies of approximately 90%, 70%, and 40% for random forest, decision tree, and Bayesian networks. Soares et al. (2019) observed that increasing the number of purpose classifications resulted in a decrease in accuracy.

Point of Interest data (POI), such as Open Street Maps and Google Places, are an increasingly rich source of features to add context and deduce trip purposes. Real-time linkage of online POIs provides more timely and accurate information about nearby locations than pre-downloaded copies of POI data, particularly for relatively dynamic object types such as shopping areas. However, the quality and availability of POI data vary by region (Ermagun et al. 2017). Online resources like OSM and Google Places API, as well as offline land use data collected by Statistics Netherlands (Centraal Bureau voor de Statistiek or CBS), are the two primary sources of information regarding POIs in this study. These datasets contain 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 than Google Places. OSM relies on crowd-sourced data and has strong community support. Additionally, OSM may be downloaded and used offline. Previous studies utilized data from Google Places API in a variety of fields, including employment data assessment (Noh et al. 2019), criminology (Vandeviver 2014), and investment information systems (Sayed et al. 2015). Land use-related research, including route planning for blind pedestrians (Cohen and Dalyot 2020), building type prediction (Atwal et al. 2022), and environmental assessment (Kloog et al. 2018), more widely applied OSM data.

This study continues research on trip purpose prediction, focusing on the utility of POI data and respondent background characteristics within the context of high-resolution location data collected in travel surveys. Travel surveys employ specific classifications of trip purpose, being interested primarily in out-of-home activities. Furthermore, travel surveys require detailed segmentation of travels in trips so that a substantial part of stops corresponds to changes in travel mode. Statistics Netherlands can link detailed background characteristics from administrative data, allowing for an evaluation of stable features versus stop-specific features.

Data from a field experiment was used to perform the study conducted in 2022 by Statistics Netherlands. Statistics Netherlands developed a smart travel survey app in 2018 (McCool et al. 2021) that expanded further in 2020 and 2021. Based on the GPS sensor data, an automated travel diary is created for the survey respondents. Each day is divided into distinct segments representing stationary periods (stops) and periods of travel (tracks). The personal movement investigation data gathered by multi-day use of the CBS app can uncover travelers' movement patterns.

This study aims to answer the general research question 'How well can we predict the trip 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 trip purposes?
2. To what extent do individual behaviors and characteristics influence the accuracy of trip purpose prediction?

2 Data

2.1 Data collection and sample sizes

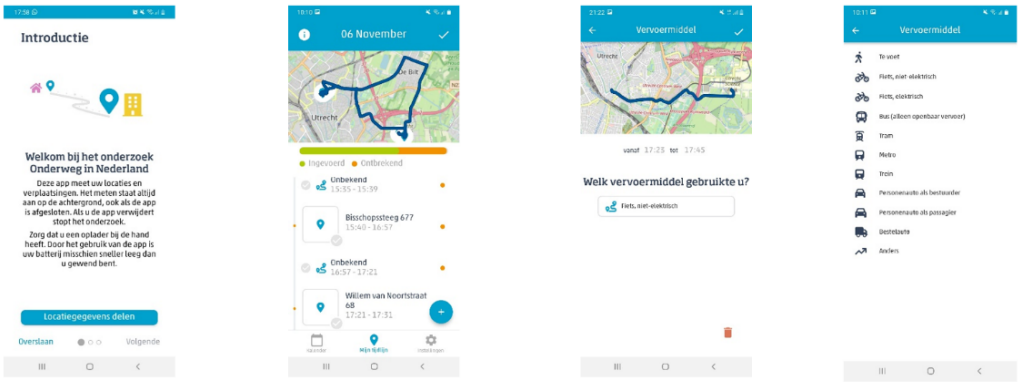
Between November 2022 and February 2023, Statistics Netherlands conducted a large-scale field test using the Statistics Netherlands Travel app ('CBS verplaatsingen' app in Dutch). Statistics Netherlands developed the cross-platform app for the regular, repeated ODiN survey (short for 'Onderweg in Nederland' or 'On your way in the Netherlands' in English). This app gave respondents several options to edit automated stop-tracks displayed during the diary reporting period. The experiment involved randomly varying three design features: the length of the reporting period (either one day or seven days), the extent of respondent editing (full editing or limited editing), and the timing of when the web diary was offered as an alternative (immediately upon invitation, at the first reminder, or the second reminder).

In the 'follow-up sample', respondents were asked to use the app for seven days and complete an online questionnaire for one day within the same week, with a version of the app allowing considerable user interaction. Respondents could choose between the app or the online questionnaire in both the 'full editing sample' and the 'limited editing sample' — the timing of when the questionnaire was offered varied experimentally. The 'full editing sample' featured an app with greater user interaction and editing capabilities than the app used in the 'limited editing sample'. The 'follow-up sample' contains 212 respondents, the 'full editing sample' 145 respondents, and the 'limited editing sample' 148 respondents. We refer to the recently published discussion paper by Schouten et al. (2024) for more details.

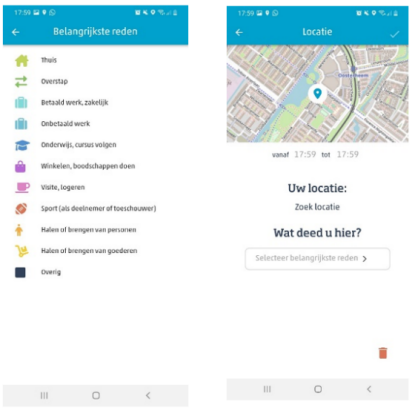
Figure 2.1 shows a preview of the app's interface. Users were asked to label travel modes and purposes. There was no automatic prediction of the labels from the app. Travel modes were classified into 11 categories (Figure 2.1a, right panel). Trip purposes were also classified into 11 categories (Figure 2.1a, left panel), which are pick-up/drop-off goods, pick-up/drop-off persons, education, sport, home, visits, paid work, unpaid work, shopping, change of travel mode, and other. Not all categories are present for every user because not every user used all categories.

Users could validate the stop and track classifications of the trips, but not all users did so (individuals validated 78.6% of identified stops and tracks). As a result, all labels show the stop or track classification. Optionally, the user was able to label the modality or purpose of a trip in the app. Not every user labeled all stops with a purpose and all trips with a transport mode. For the purposes of this study, user-labeled data is assumed to be the ground truth. However, it must be noted that stop-track segmentation is subject to error. Some tracks may actually involve multiple travel modes. In those cases, the respondents likely selected the mode that took most of the travel time. Some of the short stops may have been missed as well. Finally, the classification of working at home may have been confusing to respondents. The focus of this study is on trip purpose prediction, using the collected data from the app as training and testing data.

Figure 2.1 Trips, travel mode, and travel purpose pages in the travel app.
(a) Overview of app pages, from left to right: introduction, a preview of daily trips, the question for travel mode, and travel mode options.



(b) Overview of app pages, from left to right: question for trip purpose, location of trip purpose.



2.2 Data handling

Different kinds of location-measuring sensor data, i.e., GPS data, Wi-Fi location, and GSM data, were collected from users' phones for either one day or seven days. Sensor types differ depending on the phone type, including balanced, fused, Geolocator Platform, iOS, and network sensor. A phone can have multiple sensors with varying degrees of precision. For more details on the data quality, we refer to the recently published discussion paper by Gootzen et al. (2025).

The collected data consists of 21,397,699 observations of location and timestamps. After selecting only unique location points and eliminating location data with insufficient data quality, 12,677 stop locations from 456 users were retained. About half of the observations were classified as stops (stationary period), and the other half were tracks (moving period).

Due to only a few available observations in some categories, the trip purpose categories 'pick-up/drop-off goods' and 'pick-up/drop-off persons' that were presented separately in the app were combined into the category 'pick-up'. Similarly, 'paid work' and 'unpaid work' were also combined into the trip purpose 'work'.

The information about the distance of each track and the duration of the stop and track are reasonably important in predicting trip purposes; therefore, these two variables were added after using median smoothing on the track locations. The track's distance is defined as the Haversine great circle distance or the shortest distance between two stop points. The duration is

the difference between the end and start times in minutes. Other information about the time of day and day of the week was also added as a feature.

The start times of the stops or tracks were used to determine the day of the week, time of day, and start date. The total number of visits to regularly visited places, such as one's own house, is rather low. Visits to the same location multiple times result in slightly different locations being measured in close proximity to one another. To ensure these stops were seen as the same location, the latitude and longitude of the locations were rounded to five decimal places, resulting in an accuracy of 55 meters. The total frequency of visits at the same location, with a 55-meter margin of error, was determined. The sequence of trips and the weight of visits each day for each person were determined by using the identical trip purpose label and location, with a margin of error of 55 meters.

Administrative data is linked with the location data for respondents for whom administrative data was available. In consequence, the socio-demographic characteristics of individuals are treated as 'unknown' for missing data. We refer to Schouten et al. (2024) for information on the demographic distributions.

2.3 Data from geospatial features

A large number of attributes characterize OSM maps, describing every aspect of the platform's free tagging system. A tag is a combination of a key and a value. The community establishes unofficial criteria for the most popular tags' key and value combinations. A specific tag can comprise many keys, for example, education derived from keys amenity (various sorts of educational establishments: college, school, university), land use (a designated location for educational reasons), and office (an office space for an educational institution).

Various OSM tags were utilized, employing different tags according to the trip-purpose labels. For instance, the presence of a sports category led to the inclusion of several tags related to sports. Furthermore, tags with distinct keys that served the same purposes were grouped. The limited availability of data on some classes of trip purposes prevented the categorization of some categories. Multiple bounding boxes with varying radii were determined from OSM for different tags associated with trip purposes. Four distinct radii — 25, 35, 50, and 200 meters — were evaluated one at a time per type of POI in an exploratory way. Ultimately, one radius was selected for each type of POI. A distance of 25 meters was found to be optimal for a particular intersection or a limited portion of a street. A 35-meter distance was found to be suitable for examining a compact park or a singular city block. A radius of 50 meters was found to be appropriate for spacious crossroads, a compact group of structures, or a section of a residential area. The largest radius that was tested, 200 meters, encompasses a substantial portion of a neighborhood, a small public square, or a brief segment of a major road. A large number of features could not be solely used in the model since the variables of four distinct radii contain overlapping information. In order to maximize training time, reduce noisy attributes, and prevent over-fitting the data, the model was constrained to only one radius per tag. Table 2.1 comprises the entirety of OSM data that will be subsequently utilized as features in the model. A variable was created for each tag for all four different radii (25m, 35m, 50m, 200m), and only one radius is included in the final model based on variable importance. (2) The other category is not used in the model (only as individual tags) but was explored as a feature. An alternative operationalization is the distance to each type of POI for each location visited. This was not done for computational reasons because some types of POI are relatively rare and would require a wide-range search.

Table 2.1 Features for OSM Data. All features are binary indicators. OBJECTTYPEXX is 1 in case an object of OBJECTTYPE was found within XX meters radius and 0 otherwise.

Groups	Tags and radius
Office	offices35, schools25, commercial_land50, industrial_land35, residential_land50, cycle_way35, footway50, tertiary35, secondary25, office200, parking35, hospital200
Education	college25, education_land35, university35
Recreational	pub25, restaurant200, l50
Housing	residential_land50, lane50, cycle_way35, sidewalk25, path50, footway50, pedestrian25, tertiary35, secondary25, house50, apartment50
Health	veterinary50, pharmacy35, hospital200, doctor50, dentist25
Change of travel mode	bus_stop25, platform200, railway_land25, routes50, footway50, pedestrian25, busway50, secondary25, bus_st200
Shopping	residential_land50, retail35, shops25, cycle_way35, footway50, pedestrian25, tertiary35, secondary25, parking35, restaurant200, l50
Sport	fitness50, sport_centre50, sports35
Others	swim25, primary50, trunk25, motorway35, residential25, hotel200, police25, cinema25, parking50, bar200, aerodrome35, nursing200, schools25, all_poi50, bank25

2.4 Data Screening

2.4.1 Time-related

The analysis excluded data with insufficient quality, which was determined as users with less than one hour of data and less than 2000 location observations. Data acquired beyond the specified reference periods, which might be either one or seven days, was also omitted.

A track should ideally be located between two stops. Nevertheless, the application can identify false trip ends, such as duplicate trips, missing trips, or visits with short duration or distance (Axhausen et al. 2003; Elevelt et al. 2019). The application did not omit any data due to incorrect trip endings. As a result, certain information about the track before a stop, such as travel mode, distance, and duration, may be missing.

No outliers among the stops were removed since we assume it is reasonable to stay in a certain location, such as home, for an extended period of time. Similarly, no outliers were removed from the tracks. A lengthy journey is assumed to incorporate a break, so any travel duration beyond 24 hours was disregarded, as humans require rest after being on the road for 24 hours.

All track-related features were included to enhance stop-purpose prediction. Undefined stops and unspecified purposes (NA values) were excluded from the data. After data cleaning, the sample consists of 323 respondents with a total of 5069 stop locations.

2.4.2 Location-related

GPS measurements outside of the Netherlands' bounding box were eliminated. After eliminating around one thousand observations and applying the bounding box to OSM data, the final dataset consists of 315 respondents with a total of 4961 stop locations.

3 Methods

3.1 Machine Learning

The data will be partitioned into training and testing sets with a ratio of 80:20 from the 4961 locations available for further investigation. Due to the complexity of the data, traditional statistical approaches may not be effective for analyzing a large set of locations and geospatial features, which are considered big data. Machine learning is often regarded as one of the most effective approaches for dealing with complex data due to its efficient capacity to discover complex relationships (Zhao et al. 2020). 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).

In the training of models, the hierarchical structure of the stop-track data within respondents was ignored. Respondents visit some locations such as home or work multiple times during the reporting period. In the training, one visit to the location was selected and added to the training or test data. However, the number of times the location had been visited throughout the entire tracking period was added as feature.

This study used an Artificial Neural Network (ANN) as the initial machine learning model since ANN is able to handle such complex data, including high-dimensional data, and can adapt to changes in the input data. Some advantages of the ANNs are (1) the ability to capture nonlinearity, allowing them to identify and model complex correlations between variables that may not be easily detectable using linear regression; (2) the ability to acquire knowledge from recent data and modify their forecasts accordingly; (3) its robustness, as they are capable of effectively processing data that is noisy or incomplete without experiencing a major decline in performance; and (4) the ability to generalize from the data they are trained on to new and unknown data.

An optimal selection of features was chosen after training the ANN model, and then other machine learning models were also trained for comparison.

Based on prior research, weather data has the potential to assist trip-purpose prediction (Zhao et al. 2020). Furthermore, daily weather data from the largest and centrally located airport, Schiphol, was included in the best model. The station collects the following weather variables: daily mean wind speed, sunshine duration, daily precipitation, daily mean temperature, mean daily cloud cover, and daily mean relative atmospheric humidity. Ultimately, however, weather data features were discarded. The field experiment took place in December-January with relatively little variation in features.

All analyses were executed in Rstudio (Version 2023.12.1+402 with R 4.4.0) (R Core Team 2023), mainly using caret (Kuhn 2019) and NeuralNet packages (Günther and Fritsch 2010). All scripts are available on GitHub[1)].

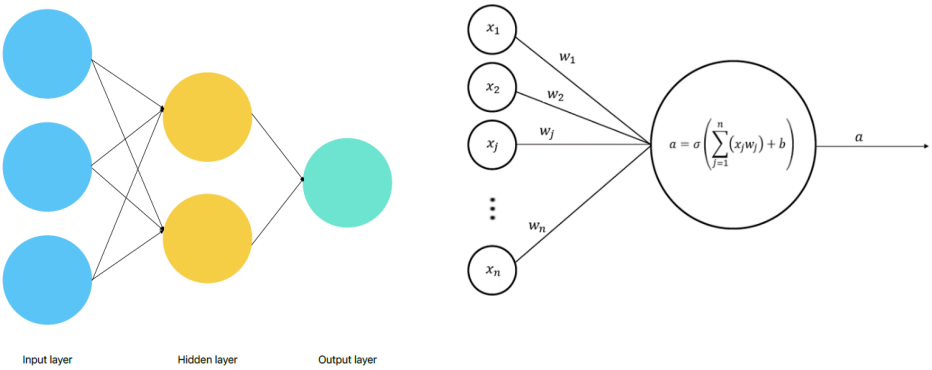
[1)] https://github.com/solichatuszhrh/travel_purpose_prediction

3.2 Artificial Neural Networks

McCulloch and Pitts (1943) proposed Artificial Neural Networks (ANN) based on the neural architecture of the human brain, which features interconnected and layered nodes. As shown in Figure 3.1, an ANN typically consists of (1) an input layer to receive data, (2) one or more hidden layers to process information through weighted connections, and (3) an output layer to produce results. Each neural network consists of a large number of processing elements with units connected through weights. In simple terms, a processing unit applies an activation function to a weighted sum of its input and passes the result as a signal to the other units (McCulloch and Pitts 1943). The network learns the weights through labeled examples. The activation function is only available in the output layer and hidden layer. The sigmoid and tanh are two commonly used activation functions. ANNs are broadly categorized into two types: feed-forward neural networks (FFNN) and recurrent neural networks (RNN). FFNN is a network type where the connections between neurons in the layers do not form a cycle. This means that the input information only moves in one direction, from the input layer to the output layer. A perceptron is a fundamental unit of a neural network with no hidden connections between its layers. On the other hand, a multilayer perceptron is a more complex neural network with a hidden layer. Introducing feedback connections to a feed-forward neural network transforms it into an RNN. The RNN, often used for time series models, is considered a network with memory due to the presence of connections between neuron layers (Kandıran and Hacınlıyan 2019).

Figure 3.1 Visualization of an ANN.

(a) ANN structure **(b) ANN with weights and bias**



The equation for a single neuron ANN, as shown in Figure 4.3 (b), can be expressed as follows

$$f(x) = \sigma\left(\sum_{j=1}^{n}(x_j w_j) + b\right) \tag{1}$$

where $f(x)$ is an output function, j is the number of inputs, σ is an activation function (reinforcing or inhibiting the signal sent by input), b is a bias, x_j is an input value, and w_j is a weight. In a more generalized form, considering multiple layers, the equation becomes a composition of these individual neuron equations across layers for the entire network.

ANNs perform exceptionally in complex tasks such as image recognition, natural language processing, and complex pattern recognition. To reduce the disparity between predicted and actual outputs, ANNs modify weights while undergoing the learning process. Furthermore, ANNs are well-suited for complex tasks involving large datasets despite the requirements of substantial computational resources and may be prone to over-fitting.

Different settings were applied for hyperparameter tuning. The hidden layer is limited to a size ranging from 1 to 5. The L2 regularization value, which is used to prevent over-fitting, is adjusted in increments of 0.01 from 0 to 0.1. The training approach employs adaptive leave-group-out cross-validation with 10 iterations, while the pre-processing involves min-max normalization.

3.3 Model comparison

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 benefits and drawbacks of each model are presented in Table 3.1.

Table 3.1 Advantages and disadvantages of machine learning models.

Model	Advantages	Disadvantages
Artificial Neural Network	nonlinearity, adaptability, generalization, scalability	computationally extensive, much training data required, black box nature, long training time
Random Forest	accuracy, robustness, interpretability, versatility	computationally extensive, less interpretable, complexity
Extreme Gradient Boosting	high performance, flexibility, regularization, efficiency	parameter tuning, complexity
Support Vector Machine	effectiveness, versatility, robustness, kernel trick	computationally extensive, memory usage, parameter tuning, less interpretable
Naive Bayes	simplicity, efficiency, scalability, performance	assumptions, probabilistic output, data sensitivity, limited model complexity

The test sets of various models were assessed using several metrics, specifically balanced accuracy and F1-score. The balanced accuracy metric distinguishes itself by giving equal weight to the correct labeling of positive and negative examples. On the other hand, the F1-score does not take into account the quantity of correctly identified negative examples or the overall number of negative examples in the dataset and focuses more on identifying positive examples, such as outliers or anomalies. Appendix A.1 provides additional details about the comparison of models and the evaluation of the metric.

4 Results

4.1 Exploratory analysis

The majority of stops occurred between the hours of 8 am. and 4 pm., coinciding with typical working hours. Conversely, there were fewer actions at night, suggesting that it is likely that most individuals stayed at home, and there were no changes in stops during this time. According to

Figure 4.1 Time distributions of the number of stops, split by time of day and day of week.

(a) Time of Day **(b) Day of Week**

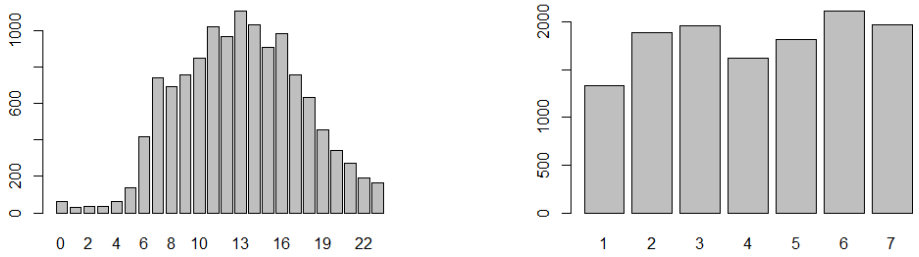


Figure 4.1, there is a high frequency of stops occurring on Fridays, while Sundays have the lowest level of activity.

Home accounted for the majority of 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%.

Typically, the journey distance via bus, bike, metro, tram, and walking was shorter (less than 20 km, excluding outliers) compared to other modes of transportation. The car had the greatest distance range of 406 km, followed by other means of transportation with a range of 279 km and trains with a range of 234 km.

Figure 4.2 displays the measurements of distance and duration for several modalities. Greater distances, conducted via various means such as cars, trains, or alternative modes of transportation, correlated with increased trip duration. Some journeys on foot and by bicycle also lasted a lengthy duration, but the distance covered was moderate. It is likely that several long excursions with short distances were due to measurement inaccuracies.

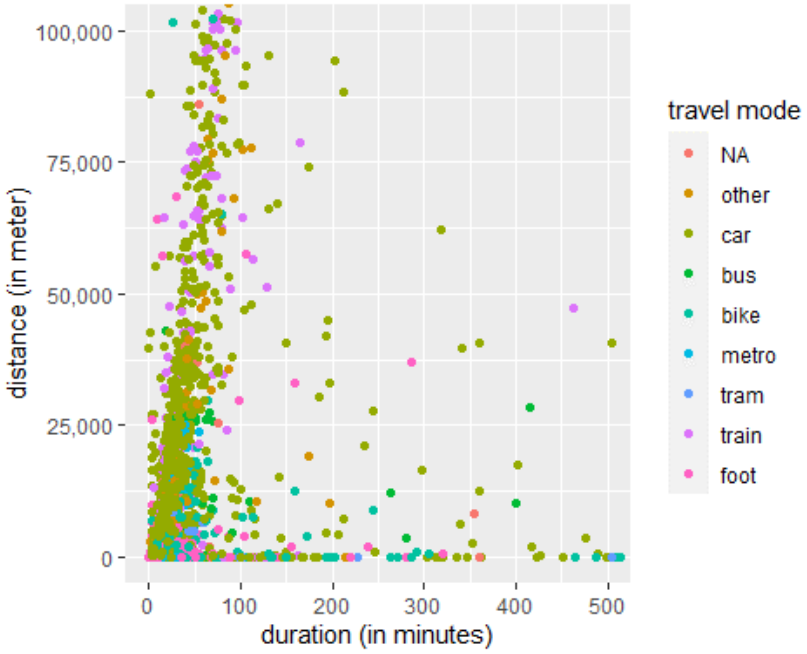
All stops had a minimum duration of less than 3 minutes, while the longest duration of all stops, except for transit, was almost 38 hours, with a transit time of 7.5 hours. Regarding the routes, bus and tram journeys had the lowest duration, lasting approximately 7 minutes. However, the tram was the only mode of transportation, with a maximum duration of 15 hours. Additional data regarding the length and distance of other modalities can be found in Appendix A.3.

4.2 Trip purpose prediction

4.2.1 ANN models

The initial analysis was performed without POI data. The training and testing data had approximately 65% balanced 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 balanced accuracy decreased by 12% when it incorporated POI data within four distinct radii. However, only the balanced accuracies for home, work, and shop are larger than 50%. The model's complexity and insufficient training data may lead to a decline in balanced accuracy. In order to reduce the duration of the training process and obtain a more efficient model, several features

Figure 4.2 Distribution of distance and duration of different Mmodality (only show trips less than 110 km and less than 525 minutes).



that frequently overlap have been eliminated. As a result, each tag will now utilize a single radius determined by its variable importance. This model enhanced the overall balanced accuracy by 10% and 15% for the training and testing sets, respectively. The test set also showed improvement in balancing accuracy, ranging from 54.9 to 89.2%.

The tags were categorized and examined in the subsequent trial 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 trip-purpose labels. Various combinations of category-only, percentage-only, category and individual tags, and percentage and individual tags were utilized. Only the model balanced accuracy of the percentage-only model decreased compared to the prior model. The balanced accuracy of the other models used in the test increased by 3–5% relative to the prior model. The model with the highest balanced accuracy in both training and testing used a combination of percentage and individual tags, yielding balanced accuracy rates of 76.7% and 72.3%, respectively. The model's balanced accuracy ranged from 62.9% (for the 'other' category) to 91.3% (for the 'home' category). Table 4.1 presents more comprehensive findings about balanced accuracy.

The ANN_6 model achieved the highest F1 scores, consistent with its balanced accuracy. When applied to a balanced dataset, the F1 score and balanced accuracy should be roughly equal. However, due to the uneven distribution of samples among the classes, the F1 score deviated significantly from the balanced accuracy, with the exception of the home class. The F1 scores for sports and visiting classes were 48.1% and 40%, respectively, whereas the balanced accuracy for these classes was 71.1% and 67.5%.

Table 4.1 Balanced accuracy of the model on the test set

Model	ANN_1	ANN_2	ANN_3	ANN_4	ANN_5	ANN_6	ANN_7	ANN_8
Overall	64.5	52.3	67.5	71.8	70.4	72.3	66.0	71.4
Pick-up/drop-off	71.8	50	72.5	69.9	67.4	75.0	72.6	77.8
Education	52.7	50	63.6	63.7	60.9	83.3	63.7	71.8
Others	58.8	50	54.9	67.5	60.7	62.9	54.2	56.8
Travel mode change	70.5	50	71.6	87.2	75.6	80.1	62.8	77.0
Sport	53.3	50	69.5	73.1	68	71.1	61.2	71.2
Home	88.8	80.2	89.2	92.5	92.4	91.3	90.2	90.5
Visit	58.5	50	60.2	71.5	64.3	67.5	58.8	66
Work	81	57.3	82.6	81.3	86.2	84.1	80.9	84.1
Shop	76.4	76.3	79.4	80.7	82.5	81.5	80.7	84.1

ANN_1: ANN model without OSM
ANN_2: ANN model with OSM data from 4 radiuses
ANN_3: ANN model with OSM data from 1 radius as count data only
ANN_4: ANN model with OSM data from 1 radius as count data and category
ANN_5: ANN model with OSM data from 1 radius as a category only
ANN_6: ANN model with OSM data from 1 radius as count data and percentage
ANN_7: ANN model with OSM data from 1 radius as a percentage only
ANN_8: Best ANN model + weather data

The ANN model that utilized OSM data, using both percentage and individual tags, achieved the highest overall balanced accuracy and was chosen as the final model. In addition, weather information was incorporated to test whether this would increase the model's performance. Including weather data did not enhance the model's performance. Overall, the balanced accuracy of nearly all classes decreased when weather data was included. The balanced accuracy of the pickup, sport, and shopping classes showed a slight increase, possibly due to these classes being outdoor activities. The final configuration for the ANN was a model consisting of five hidden layers. To prevent over-fitting, a regularization parameter of 0.06 was employed (Figure 4.3 (a)). The final model for model comparison utilized the settings and data from the best ANN model, specifically ANN_6, without incorporating weather data.

4.2.2 Comparison of different models

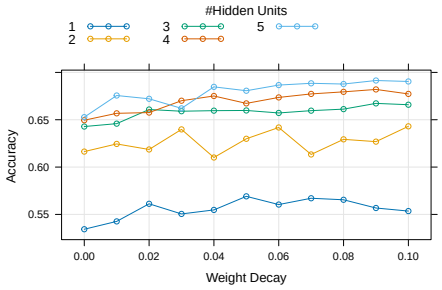
The RF, XGB, SVM, and NB models were trained using the same set of training control parameters and data sets. The optimal hyperparameter configuration for the final RF model consisted of five selected random predictors. For XGB, the final model included 19 trees with a maximum depth of 6, a learning rate of 0.3, a regularization parameter of 0, and a minimum leaf weight of 1. The SVM model showed no effects of the cost parameter on the accuracy. The final SVM model had a cost function value of 8. The NB model showed no effects of the Laplace correction, which was set to 0.02. The final NB had a bandwidth value of 0.7.

The RF model's balanced accuracy was perfect for the training set but not for the test set. The XGB model also achieved near-perfect training balanced 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. Figures 4.3 (b) and 4.3 (c) show the sensitivity in balanced accuracy for the RF and XGB training processes. Note, the y-axes for these figures do not start at 0.

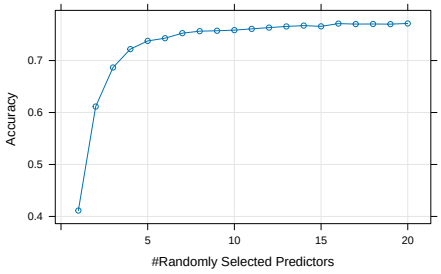
The SVM model achieved a similar balanced accuracy compared to the ANN (Figure 4.3 (d)). Both models achieved a balanced accuracy of approximately 70%. The NB model performed worse than the other models. The classifier was unable to differentiate between classes and produced

Figure 4.3 Balanced accuracy of machine learning models (based on test set).

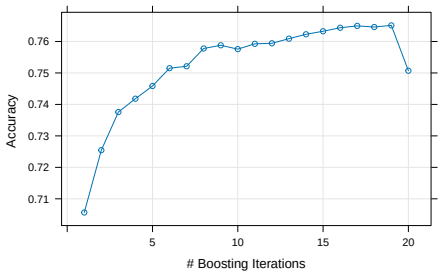
(a) ANN performance



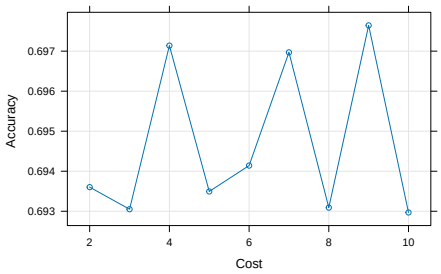
(b) RF performance



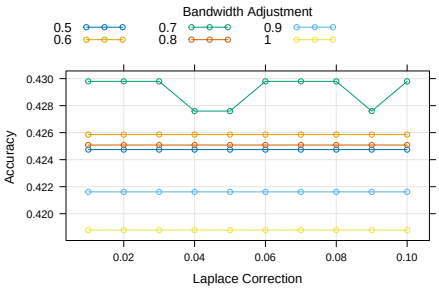
(c) XGB performance



(d) SVM performance



(e) NB performance



almost random predictions, leading to a balanced accuracy rate of approximately 43%. The over-representation of the home class may have led to the classifier projecting nearly every class as home. Figure 4.3 (e) illustrates the performance of the models.

After considering the training duration and balanced accuracy, we found that XGB was the best among the models, achieving a balanced accuracy of 96.7% and 77.7% for the train and test sets, respectively.

Table 4.2 Balanced accuracy of the model for the test set.

Model	ANN	RF	XGB	SVM	NB
Overall	72.3	77.5	77.7	70.8	42.7
Pick-up/drop-off	75.0	78.2	75.3	70.7	50.0
Education	83.3	80.3	83.1	80.4	50.0
Others	62.9	69.0	74.1	64.4	50.0
Travel mode change	80.1	79.1	83.3	71.7	50.0
Sport	71.1	73.1	81.2	68.1	50.0
Home	91.3	92.8	93.2	90.7	52.6
Visit	67.5	62.3	74.6	67.3	50.0
Work	84.1	90.8	87.9	83.9	49.9
Shop	81.5	86.0	83.8	80.1	55.5

The ANN_6 model was selected as the best ANN model and will, from here on, be referred to as ‘the ANN model’. The other models represent the best models as identified in Figure 4.3.

An ANN performs better when it has more hidden layers, but it has not yet reached the highest level of balanced accuracy. When the RF model incorporated more variables, it became more accurate and reached stability in five variables. The XGB model indicated similar inflexibility to the RF model. Similar to the ANN, the SVM failed to reach robust, balanced accuracy when setting the cost function’s maximum value to 10. Unfortunately, the NB did not show any modifications except for the 0.8 and 0.9 bandwidth adjustments. Table 4.2 displays the balanced accuracy of multiple models.

4.2.3 Feature selection

When adding information from OSM, no socio-demographic features were evaluated as important to the model anymore. Without the addition of OSM variables, age was one of the most important features in predicting trip purpose. After experimenting with various combinations of included variables, it became apparent that variables indicating past visits to a location played a crucial role in adding POI data. Because of the design of the study, variables on past visits were only available for at most seven days of data collection. Hence, predictions based on seven days of data were superior to predictions based on data for one day. Appendix A.4 provides additional information regarding the variable importance.

XGB identified the top five important factors as the duration of stops, the overall frequency of visits, the number of shops within a 25-meter radius, the time of day, and the weight of visits with the same label per person. The ANN model prioritized the number of sports venues over shops, rendering the time of day irrelevant.

4.2.4 Confusion matrix

Tables 4.3 and 4.4 display the confusion matrix of the best ANN and XGB models, revealing the misclassification of each class. The models often classified visits as home. The reason was that

individuals were engaging in physical activities within their homes, and it was also possible that people were visiting others in their homes. The 'other' class was incorrectly labeled as either visit or pick-up/drop-off. The 'other' class's balanced accuracy might increase due to the ambiguity around the activities chosen by individuals who selected these options. The home class exhibited low misclassification rates due to the large sample size, allowing it to continue learning until the model reached its optimal state. The identical condition was likewise imposed on the work and shop class.

The class-specific precision and recall, as well as the macro-average precision and recall, are also shown. The ANN model achieves moderate precision and recall across most classes, with stronger performance in dominant classes like home and work, where recall and precision are relatively high. Classes, such as 'other' and 'visit', show low precision, indicating frequent misclassifications. The class 'education' has a high recall, meaning almost all actual education stop purposes are captured, but with moderate precision, showing some misclassifications with other classes like 'work.' The ANN model has good overall balance, especially in high-frequency classes, but struggles with precision in minor or ambiguous classes. The XGB model outperforms the ANN in precision and recall, especially in high-frequency classes.

Table 4.3 Confusion matrix of the best ANN model

				Predicted						
Observed	Pick-up/ drop-off	Education	Other	Other trvl mode	Sport	Home	Visit	Work	Shop	Total
Pick-up/drop-off	41	0	4	0	1	5	4	3	7	65
Education	0	12	0	0	0	0	0	1	0	13
Other	9	1	22	3	3	3	7	4	10	62
Other trvl mode	2	0	1	21	0	0	1	8	3	36
Sport	3	0	3	0	13	0	1	4	0	24
Home	4	2	8	2	2	377	15	14	12	436
Visit	3	1	10	0	3	11	24	2	4	58
Work	10	1	6	5	4	7	3	116	4	156
Shop	6	1	19	3	4	3	7	7	88	138
Total	78	18	73	34	30	406	62	159	128	988
Precision	0.5256	0.6667	0.3014	0.6176	0.4333	0.9286	0.3871	0.7296	0.6875	0.5864
Recall	0.6308	0.9231	0.3548	0.5833	0.5417	0.8647	0.4138	0.7436	0.6377	0.6326

Table 4.4 Confusion matrix of the best XGB model

				Predicted						
Observed	Pick-up/ drop-off	Education	Other	Other trvl mode	Sport	Home	Visit	Work	Shop	Total
Pick-up/drop-off	42	0	4	1	1	5	4	6	8	65
Education	0	12	0	0	2	1	0	2	0	13
Other	8	1	38	1	3	4	5	3	11	62
Other trvl mode	1	0	2	23	0	0	0	1	5	36
Sport	6	0	2	0	19	0	0	0	0	24
Home	3	1	3	2	2	384	17	13	6	436
Visit	3	1	7	0	1	5	32	2	4	58
Work	6	2	7	5	1	5	4	127	3	156
Shop	9	1	10	2	1	2	0	5	91	138
Total	78	18	73	34	30	406	62	159	128	988
Precision	0.5385	0.6667	0.5205	0.6765	0.6333	0.9457	0.5161	0.7987	0.7109	0.6674
Recall	0.6462	0.9231	0.6129	0.6389	0.7917	0.8807	0.5517	0.8141	0.6594	0.7243

5 Discussion

In this study, trip-purpose prediction models were trained on data collected by a smartphone app. The best model's balanced accuracy is a bit lower than in past studies, but it is able to classify a greater number of trip-purpose labels than in previous research. Increasing the number of labels complicates the process of making deductions and predictions, while having imbalanced data across different classes negatively impacts the model's performance.

With the amount of available training data, the model is based on an ANN with five hidden layers. In the future, adding layers to the training process can lead to longer training times, but it also has the potential to improve the model's balanced accuracy and this would require a larger amount of training data.

The model achieves satisfactory balanced accuracy on the initial attempt when trained without OSM data, and some sociodemographic factors are essential. When OSM data is included, only the background variable age is significant in the models. The age variable accounts for 18% explainability in the ANN model and 10% in the XGB model. According to the variable importance study, the most important variable is the frequency of visiting the same label on a daily basis, with a significance value of 30%. The overall visit frequency to the same area is crucial, representing 45% in the ANN model and 25% in the XGB model. This demonstrates that a data collection period of multiple days is preferable to a period of one day. It demonstrates the variety in an individual's day. The study concludes that individual behaviors are more accurate predictors of trip purpose background variables (administrative data).

A single location might serve multiple trip purposes, with possibly different stop durations. This suggests that we can use temporal patterns to identify the purpose of the trip. The inclusion of OSM-based features increased the model's performance. The OSM data was downloaded once and used offline, which results in the connotation that changes over time in OSM's dynamic database were not included in the study. For example, a shop's functionality may change to a sports facility. In OSM, the number of recreational facilities is important data. The presence of a higher number of shops and sports facilities within a 25-meter radius significantly enhances the probability of making a stop.

The inclusion of weather data did not enhance the model performance. One potential reason for this is that the weather variance was restricted due to the data being collected during the winter season. The weather data for most trip objectives are consistent, as shown in Appendix A.5, due to the combination of outdoor and indoor activities. The model exhibits a marginal improvement compared to a model that does not use weather data for pickup, sports, and shop classes. In these categories, the dominance of outside activities accounts for the observed improvement. Because the duration of stops does not vary and is often short, the transit class balanced accuracy does not improve. When evaluating each activity individually, it is important to consider indoor and outdoor activities because lower temperatures and shorter daylight hours during the winter can increase indoor activity.

The confusion matrix indicates that despite the accurate classification of classes, the limited number of observations, notably in sports and education, results in low-balanced accuracy. Among all classes, the 'other' class exhibits the lowest balanced accuracy and F1 score. Several classes are more likely to be labeled inconsistently, such as work-from-home activities as either work or home. Additionally, there may be ambiguity between labels such as shopping and pickup. In future studies, administrative variables, such as the respondents' profession, could

potentially help to distinguish between these labels. From the confusion matrices, it was observed that the ANN model often incorrectly predicts 'pickup/drop-off' as a work, while the XGB misclassified it as a shop.

The inclusion of spatiotemporal patterns and OSM data accounts for the absence of sociodemographic characteristics. The significance of sociodemographic characteristics in the model's importance is evident when OSM variables are not included.

Under the condition that parameters remain unchanged, such as the same season and a number of trip purpose labels, the prediction model is reasonably accurate, particularly for the home class. Since the data was collected during winter only, it is possible that this model may not be applicable to data collected in other seasons. Forecasting the 'other' class is challenging due to its extensive scope.

Obviously, it cannot be ruled out that respondents made errors in labeling the purposes of stops. Such errors may be the consequence of satisficing behavior where respondents minimize effort. However, labeling errors may also occur because of misunderstanding definitions. It is likely that some purposes are less sensitive to such errors. However, even for the most frequently visited location, home, working from home may be labeled differently by different respondents. Labeling errors will reduce the maximal performance.

Two different models, ANN and XGB, perform relatively similarly but with very different feature importances. This implies some caution is needed with substantive interpretations of the features. Features are mutually associated. Given that different features appear in different models, ensemble approaches may give some improvement over single approaches.

6 Conclusion

In conclusion, the performance of the travel purpose prediction models trained in this study suggests that it is worth further developing such models for app-based travel surveys. During the initial trial, the optimal model for a residential classification achieved a balanced accuracy of over 80% for the home class on the testing set and an overall balanced accuracy of 60%, which is considered satisfactory given that just GPS data and sociodemographic factors were utilized. The model's balanced accuracy increased to more than 90% for the home class and over 70% for the overall model by incorporating multiple point-of-interest features based on OSM data. While some significant variables were derived from OSM data, most were acquired using temporal data. Based on the results of this study, sociodemographic features do not appear to be essential predictors of trip purpose when spatiotemporal information is available.

Recording spatial and temporal patterns diminishes the importance of certain characteristics. This is a promising indication of accurately predicting a trip's purpose in real-time. The absence of a request for the users' sociodemographic characteristics prevents immediate access to this information in real time. Nevertheless, people may visit the same location several times on different occasions. In order to accurately determine the purpose of the trip, it is important to include additional information, such as the respondents' occupation. Overall, using sensor data from a smartphone-installed travel diary application successfully predicts the purpose of a trip. Utilizing spatial and temporal patterns is valuable for predicting trip purposes, and individual

behaviors have minimal influence on the balanced accuracy of trip-purpose prediction. Some types of stops would be more important for making accurate predictions than others.

This study is subject to a number of limitations and simplifications. Firstly, the data collection period was limited to one season. For example, weather data features showed little variation. Secondly, only one observation was selected per location. The nested structure of stops within respondents was largely ignored. However, the number of visits to the location during the tracking period was added as a feature. A multi-level/hierarchical structure may improve classification performance. Thirdly, indoor and outdoor activities were intermingled, and sample sizes per class were imbalanced. Additionally, certain information, such as respondents' occupation, was excluded. Fourthly, in the training and testing, no distinction was made between validation data sets and test data sets. Also, settings were not tuned using the validation data set, which may have resulted in overfitting. Lastly, individual characteristics, such as the tendency for teachers to visit educational places more frequently than others, were not further explored. Hence, the model may be improved by opting for better tags in OSM, leveraging the most up-to-date OSM data, and collecting data from multiple seasons with a larger and more balanced sample. An increase in the quantity of data points is 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.

There are also a few open areas of research. This study did not perform an external validation, i.e., an application of the trained models to location data collected outside the field test. This is an imperative step as location data properties such as frequency are known to change over time (e.g., new versions of operating systems and app store policies). In a follow-up study, external validation is performed. This also leads to the most important open area: the role of the respondent. In the app-assisted travel survey, respondents will be actively involved regardless of stop prediction accuracy. They are crucial in checking and adjusting stop-track segmentation. If a respondent makes an edit, then stop labels need to be added. However, from a transparency and legal point of view, respondents should also be able to see and check. Given that predictions will never be perfect, there is a trade-off between training of models and respondent involvement. This trade-off has a large impact on the user interface and user experience. Another open research area follows from these considerations, namely the extent to which respondents are able and motivated to correctly label stops. There likely is a respondent-dependent limit to how far errors can be avoided. The implications for predictions are not yet fully understood. Finally, an open area that was only implicitly addressed is the study duration. From a prediction viewpoint, multiple days are clearly preferable. How many days is ideal is, however, yet unclear. Again, respondent motivation will play a decisive role. Follow-up research should address the study duration and the role of the respondent.

Acknowledgments

The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands.

The authors would like to thank Joep Burger for careful reading and providing useful comments on a former draft of this manuscript.

References

Atwal, K., T. Anderson, D. Pfoser, and A. Züfle (2022). "Predicting building types using OpenStreetMap". In: *Scientific Reports* 12.19976, pp. 1–13. DOI: 10.1038/s41598-022-24263-w.

Axhausen, K., Schönfelder, W. S., M. J. Oliveira, and U. Samaga (2003). "80 weeks of GPS-traces: Approaches to enriching the trip information". In: *Arbeitsberichte Verkehrs- und Raumplanung* 178. DOI: 10.3929/ethz-a-004570614.

Bohte, W., M. K., and W. Quak (2008). "A Method for Deriving Trip Destinations and Modes for GPS-based Travel Surveys". In: *Research in Urbanism Series* 1, pp. 127–143. DOI: 10.7480/rius.1.201.

Brodersen, K., C. Ong, K. Stephan, and J. Buhmann (n.d.). "The Balanced Accuracy and Its Posterior Distribution". In: *2010 20th International Conference on Pattern Recognition*. IEEE, pp. 23–26. DOI: 10.1109/ICPR.2010.764.

Calabrese, F., M. Diao, G. Di Lorenzo, J. Ferreira, and C. Ratti (2013). "Understanding individual mobility patterns from urban sensing data: A mobile phone trace example". In: *Transportation Research Part C: Emerging Technologies* 26, pp. 301–313. DOI: 10.1016/j.trc.2012.09.009.

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

Cohen, A. and S. Dalyot (2020). "Route planning for blind pedestrians using OpenStreetMap". In: *Environment and Planning B: Urban Analytics and City Science* 48.6, pp. 1511–1526. DOI: 10.1177/2399808320933907.

Cui, Y., C. Meng, Q. He, and J. Gao (2018). "Forecasting current and next trip purpose with social media data and Google Places". In: *Transportation Research Part C: Emerging Technologies* 97, pp. 159–174. DOI: 10.1016/j.trc.2018.10.017.

Elevelt, A., W. Bernasco, P. Lugtig, S. Ruiter, and V. Toepoel (2019). "Where You at? Using GPS Locations in an Electronic Time Use Diary Study to Derive Functional Locations". In: *Social Science Computer Review* 39.4, pp. 509–526. DOI: 10.1177/0894439319877872.

Ermagun, A., Y. Fan, J. Wolfson, G. Adomavicius, and K. Das (2017). "Real-time trip purpose prediction using online location-based search and discovery services". In: *Transportation Research Part C: Emerging Technologies* 77, pp. 96–112. DOI: 10.1016/j.trc.2017.01.020.

Feng, T. and H. Timmermans (2015). "Detecting activity type from GPS traces using spatial and temporal information". In: *European Journal of Transport and Infrastructure Research* 15.4. DOI: 10.18757/ejtir.2015.15.4.3103.

Gootzen, Y., J. Klingwort, and B. Schouten (2025). "Data quality aspects for location-tracking in smart travel and mobility surveys". In: *Discussion paper, Statistics Netherlands*.

Grandini, M., E. Bagli, and G. Visani (2020). "Metrics for Multi-Class Classification: an Overview". In: *ArXiv e-prints*. DOI: 10.48550/arXiv.2008.05756. eprint: 2008.05756.

Günther, F. and S. Fritsch (2010). "neuralnet: Training of Neural Networks". In: *The R Journal* 2.1, pp. 30–38. DOI: 10.32614/RJ-2010-006.

Jin, Z., Y. Chen, C. Li, and Z. Jin (2022). "Trip Destination Prediction Based on Hidden Markov Model for Multi-Day Global Positioning System Travel Surveys". In: *Transportation Research Record* 2677.2, pp. 577–587. DOI: 10.1177/03611981221107919.

Kakar, A. (2020). "Trip Purpose and Prediction". In: *International Journal of Engineering Research & Technology* 9.10. DOI: 10.17577/IJERTV9IS100143.

Kandıran, E. and A. Hacınlıyan (2019). "Comparison of Feedforward and Recurrent Neural Network in Forecasting Chaotic Dynamical System". In: *AJIT-e: Academic Journal of Information Technology* 10.37, pp. 31–44. DOI: 10.5824/1309-1581.2019.2.002.x.

Kloog, I., L. Kaufman, and K. De Hoogh (2018). “Using Open Street Map Data in Environmental Exposure Assessment Studies: Eastern Massachusetts, Bern Region, and South Israel as a Case Study”. In: *International Journal of Environmental Research and Public Health* 15.11, p. 2443. DOI: 10.3390/ijerph15112443.

Kuhn, M. (2019). *The caret Package*. URL: https://topepo.github.io/caret.

Kusakabe, T. and Y. Asakura (2014). “Behavioural data mining of transit smart card data: A data fusion approach”. In: *Transportation Research Part C: Emerging Technologies* 46, pp. 179–191. DOI: 10.1016/j.trc.2014.05.012.

Liao, C., C. Chen, S. Guo, Z. Wang, Y. Liu, K. Xu, and D. Zhang (2022). “Wheels Know Why You Travel: Predicting Trip Purpose via a Dual-Attention Graph Embedding Network”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6.1, pp. 1–22. DOI: 10.1145/3517239.

McCool, D., P. Lugtig, O. Mussmann, and B. Schouten (2021). “An App-Assisted Travel Survey in Official Statistics: Possibilities and Challenges”. In: *Journal of Official Statistics* 37.1, pp. 149–170. DOI: 10.2478/jos-2021-0007.

McCulloch, W. and W. Pitts (1943). “A logical calculus of the ideas immanent in nervous activity”. In: *bulletin of mathematical biophysics* 5.4, pp. 115–133. DOI: 10.1007/BF02478259.

Nguyen, M., J. Armoogum, J. Madre, and C. Garcia (2020). “Reviewing trip purpose imputation in GPS-based travel surveys”. In: *Journal of Traffic and Transportation Engineering (English Edition)* 7.4, pp. 395–412. DOI: 10.1016/j.jtte.2020.05.004.

Noh, H., E. Kramer, and A. Sun (2019). “Development of Strategic Regional Employment Data Assessment using Google Places API”. In: *Transportation Research Record* 2673.11, pp. 254–263. DOI: 10.1177/0361198119852068.

Oliveira, M., P. Vovsha, J., and M. Mitchell (2014). “Evaluation of Two Methods for Identifying Trip Purpose in GPS-Based Household Travel Surveys”. In: *Transportation Research Record* 2405.1, pp. 33–41. DOI: 10.3141/2405-05.

Omrani, H. (2015). “Predicting Travel Mode of Individuals by Machine Learning”. In: *Transportation Research Procedia* 10, pp. 840–849. DOI: 10.1016/j.trpro.2015.09.037.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Sasaki, Y. (2007). “The truth of the F-measure”. In: *Teach Tutor Mater*. URL: https://www.researchgate.net/publication/268185911_The_truth_of_the_F-measure.

Sayed, S., E. Zidan, and R. Ibrahim (2015). “Implementation of an Investment Information System Based on Google Maps API”. In: *International Journal of Advanced Research in Computer and Communication Engineering* 4.

Schouten, B., D. Remmerswaal, A. Elevelt, J. K. de Groot, T. J. Schijvenaars, M. Schulte, and M. Vollebregt (2024). “A smart travel survey: Results of a push-to-smart field experiment in the Netherlands”. In: *Discussion paper, Statistics Netherlands*.

Shafique, M. and E. Hato (2016). “Travel Mode Detection with Varying Smartphone Data Collection Frequencies”. In: *Sensors* 16.5, p. 716. DOI: 10.3390/s16050716.

Soares, E., K. Revoredo, B. F., C. A. d. M. Quintella, and C. Campos (2019). “A Combined Solution for Real-Time Travel Mode Detection and Trip Purpose Prediction”. In: *IEEE Transactions on Intelligent Transportation Systems* 20.12, pp. 4655–4664. DOI: 10.1109/TITS.2019.2905601.

Stopher, P., Q. Jiang, and C. FitzGerald (2005). “Processing GPS data from travel surveys”. In: *28th Australasian Transport Research Forum, ATRF 05*. URL: https://asu.elsevierpure.com/en/publications/processing-gps-data-from-travel-surveys.

Vandeviver, C. (2014). “Applying Google Maps and Google Street View in criminological research”. In: *Crime Science* 3.1, pp. 1–16. DOI: 10.1186/s40163-014-0013-2.

Wang, F. and C. Ross (2018). “Machine Learning Travel Mode Choices: Comparing the Performance of an Extreme Gradient Boosting Model with a Multinomial Logit Model”. In: *Transportation Research Record* 2672.47, pp. 35–45. DOI: 10.1177/0361198118773556.

Xiao, G., Z. Juan, and C. Zhang (2016). “Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization”. In: *Transportation Research Part C: Emerging Technologies* 71, pp. 447–463. DOI: 10.1016/j.trc.2016.08.008.

Xiao, Y., D. Low, T. Bandara, P. Pathak, H. Lim, and D. Goyal (2012). “Transportation activity analysis using smartphones”. In: *2012 IEEE Consumer Communications and Networking Conference (CCNC)*. IEEE, pp. 14–17. DOI: 10.1109/CCNC.2012.6181051.

Zhao, X., X. Yan, A. Yu, and P. Van Hentenryck (2020). “Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models”. In: *Travel Behaviour and Society* 20, pp. 22–35. DOI: 10.1016/j.tbs.2020.02.003.

Appendices

A Model comparison and evaluation metrics

A.1 Model comparison

1. Random Forest

RF unites bootstrapping and random feature selection where every single tree in bootstrapping escalates with a various training set that is randomly selected with sampling with replacement method. Consequently, some observations could be selected more than once, and some others will be omitted, which is known as out-of-bag (OOB) observations. Random selection of attributes is combined into bootstrapping to evade correlation between trees, and at each splitting mode, it allocates merely a random subgroup of variables (Cheng et al. 2019). Using the Gini index, one can ascertain how nodes on a decision tree branch are classified.

2. Extreme Gradient Boosting

XGB is a tree-based ensemble approach based on additive training and iteratively grows low-depth decision trees to minimize a specified loss function. The estimation process, however, constantly assigns greater weight to the situations that are incorrectly predicted by earlier trees. The combined outcomes of all the created trees define the final model output (Wang and Ross 2018).

3. Support Vector Machine

SVM forms a hyperplane (linear and nonlinear) in a high-dimensional area to accomplish the largest split between classes. To separate classes in the nonlinear cases, the kernel function $(\phi(x_i))$ is used, in which the radial basis function (RBF) is the most well-known one. RBF is described as $K(x_i, x_j) = exp\ (-\gamma||x_i - x_j||^2)$, where γ is a kernel function and x_i, x_j are input data. An optimal hyperplane must be obtained; thus, the maximization of the margin between linear decision boundaries is achieved.

4. Naïve Bayes

NB estimates the missing features of data using the naïve Bayes probabilistic model, calculates the class probability, and selects the highest probability. It handles all variables as discrete to prevent postulating the distribution form of $p(c|F)$. It might be, however, that the number of data is inadequate to label the relation of all observed variable combinations. It is also presuming each element F is independent conditionally of other Fs when c is given (Kusakabe and Asakura 2014).

A.2 Model evaluation

To define the best model for trip purpose prediction, numerous model evaluation metrics were handled to get a more universal viewpoint of the model's performance (Grandini et al. 2020). Accuracy is determined for the overall model, and to evaluate the model in each class, balanced accuracy, and F1-score are defined.

1. Accuracy and Balanced Accuracy

The most renowned metric in multi-class classification, accuracy, is computed from the confusion matrix. Accuracy measures how much the model correctly predicts the complete dataset. Balanced accuracy is the average value of recalls in each class. Accuracy and balanced accuracy are inclined to converge to the same value when the data is not imbalanced. However, for an imbalanced class, balanced accuracy is more insensitive than accuracy (Brodersen et al. n.d.).

$$\text{Balanced Accuracy} = \frac{\frac{\text{True positive}}{\text{Total row}_1} + \frac{\text{True negative}}{\text{Total row}_2}}{2}$$

2. F1-Score

One of the performance assessments of the classification model is F1-Score. It is a combination of precision and recall measures using the harmonic mean concept and can be translated as a weighted average of these measures. Recall aims to maximize the number of true positives, whereas precision aims to minimize the number of false positives. Smaller classes will have more weight, and it compensates models with similar precision and recall values (Sasaki 2007).

$$\text{F1} = \left(\frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} \right)$$

A.3 Figures of travel distances and durations

Figure A.1 Duration (x-scale) vs. distance (y-scale). Color indicates travel mode.

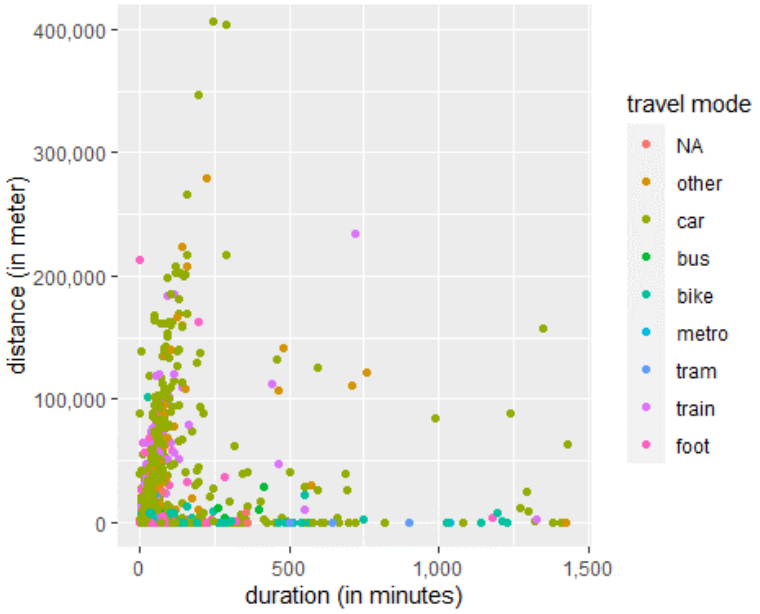


Figure A.2 Distance (y-scale) per travel mode (x-scale).

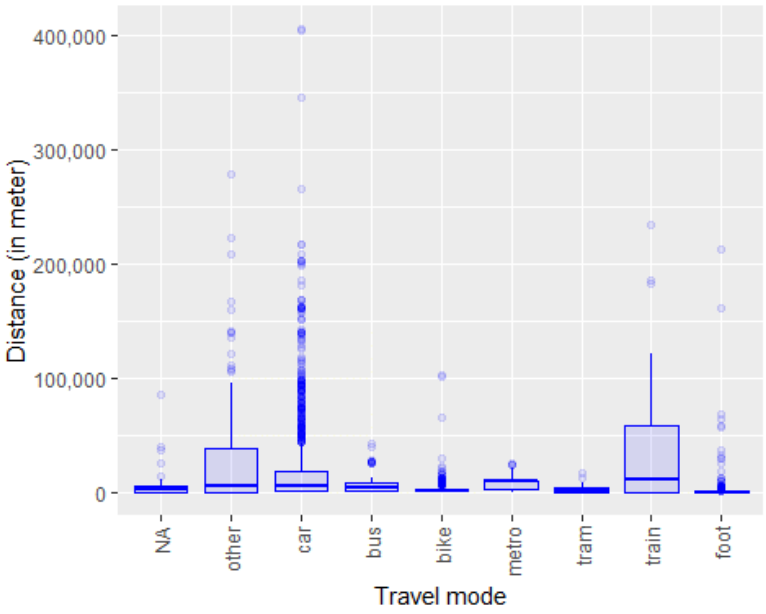


Figure A.3 Duration (y-scale) per travel mode (x-scale).

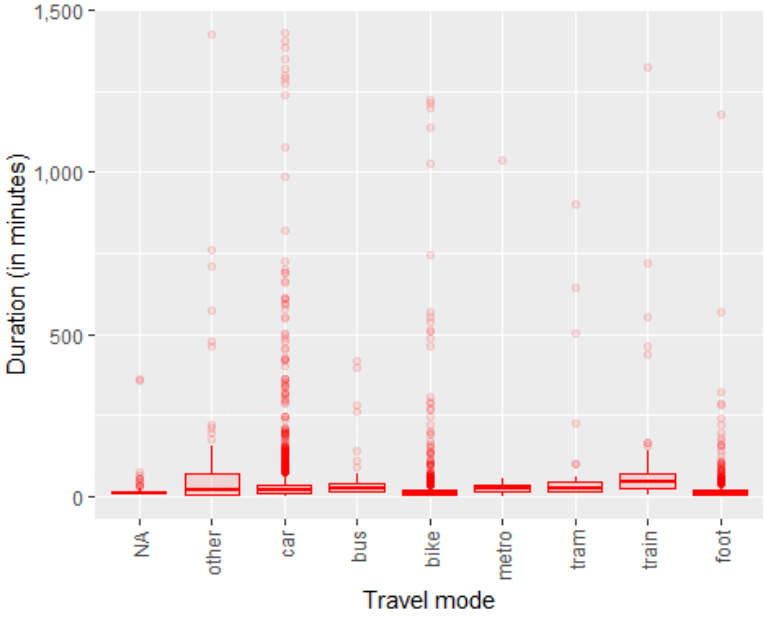
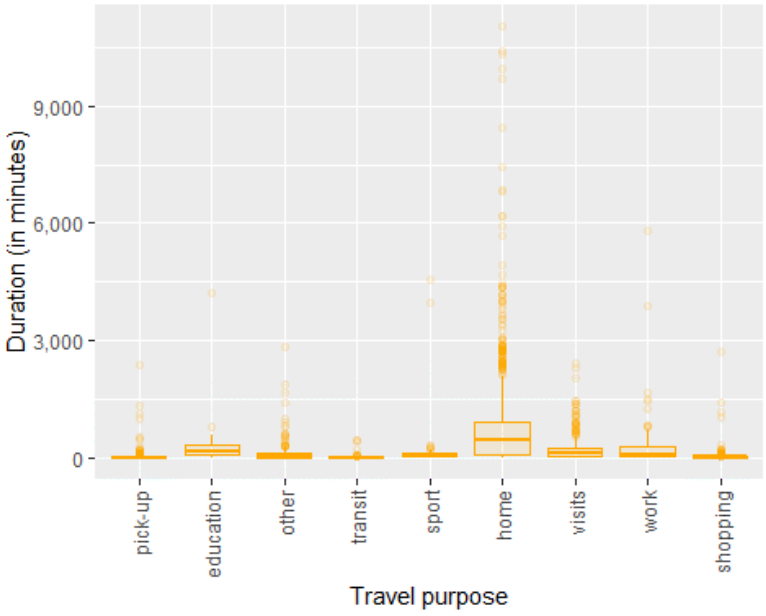


Figure A.4 Duration (y-scale) per trip purpose (x-scale).



A.4 Variable importance

These results derived by the `varImp` function from the `caret` package in `R`. This function calculates the variable importance with specific methods depending on the model use.

Table A.1 Variable importance of the best ANN model.

Variable	Importance
stops duration	100.00
weight of visit for the same purpose (per person per day)	29.96
total frequency of visits to the same location	29.88
number of sports tag within 25m	22.43
number of swim tag within 50m	21.83
number of shops tag within 25m	20.65
sequence of visits to the same location (per person)	19.76
age of respondents	18.39
number of college tag within 25m	17.62
sequence of visits for the same purpose (per person)	17.42
number of office tag within 35m	17.37
number of university tag within 50m	17.20
number of house tag within 25m	16.03
number of all POIs tag within 50m	15.84
number of aerodrome tag within 25m	15.46
number of railway land tag within 35m	15.16
number of industrial land tag within 200m	14.87
travel mode	14.80
number of cycleway tag within 25m	14.41
number of sport center tag within 35m	13.63

Table A.2 Variable importance of the best XGB model.

Variable	Importance
stops duration	100.000
total frequency of visits to the same location	80.430
number of shops tag within 25m	35.693
time of day	22.686
weight of visit for the same purpose (per person per day)	22.170
number of house tag within 25m	20.920
next stop (work)	18.034
number of routes tag within 200m	15.842
sequence of visits for the same purpose (per person)	15.242
duration of the previous track	15.098
distance of the previous track	14.860
previous stop (work)	11.162
age of respondents	10.891
latitude	10.434
longitude	8.266
number of all POIs tag within 50m	7.449
number of sports tag within 25m	6.359
start date of participation	6.165
weight of visit for the same location (per person per day)	6.029
household income	5.569

A.5 Figures of weather data and trip purpose

Figure A.5 Daily mean wind speed (y-scale) per trip purpose (x-scale).

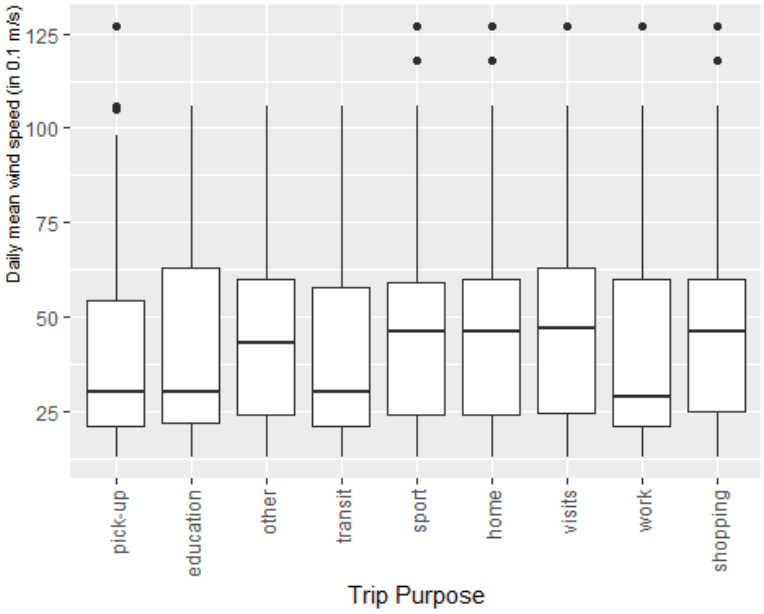


Figure A.6 Daily precipitation amount (y-scale) per trip purpose (x-scale).

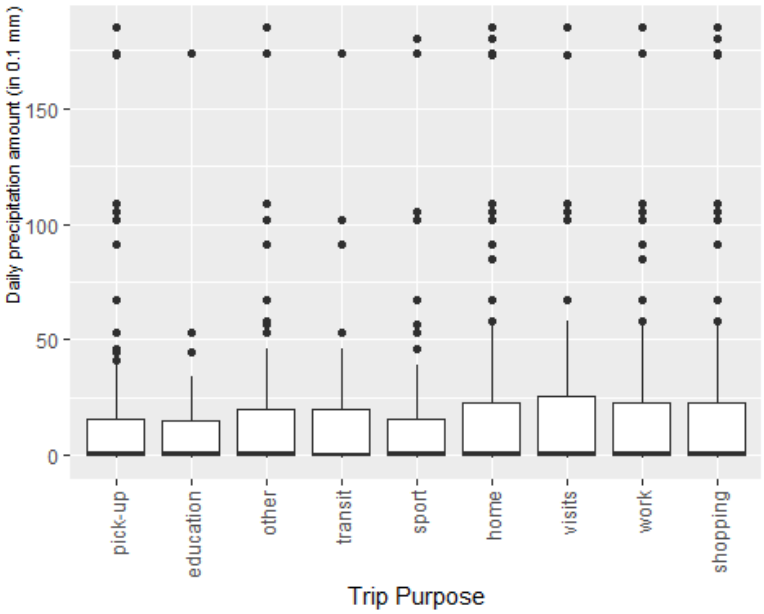


Figure A.7 Mean daily cloud cover (y-scale) per trip purpose (x-scale).

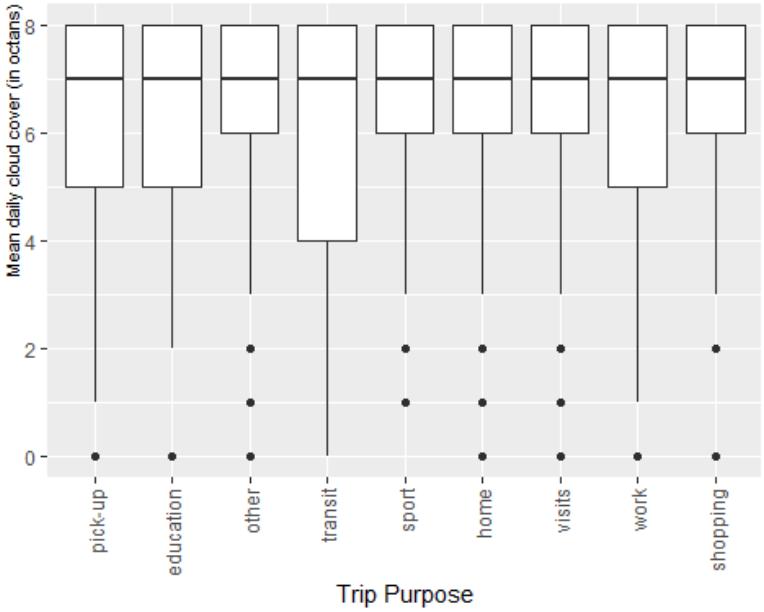


Figure A.8 Daily mean sunshine duration (y-scale) per trip purpose (x-scale).

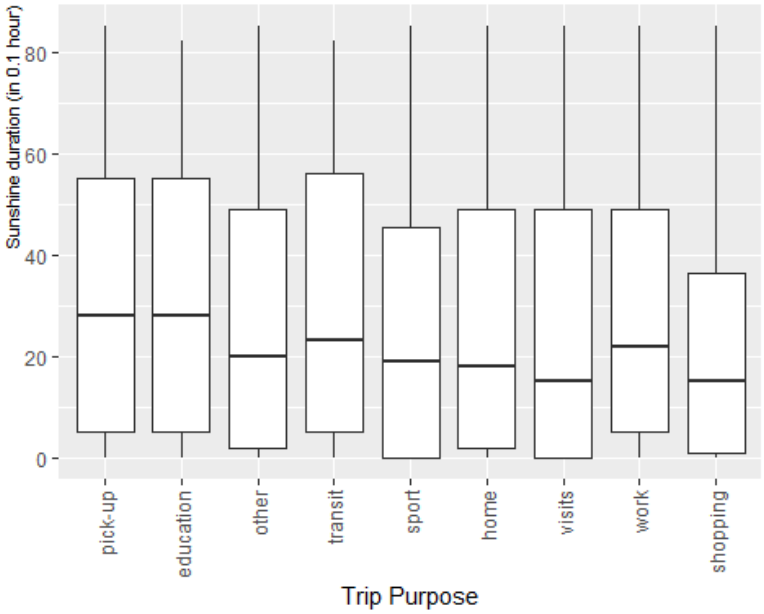


Figure A.9 Daily mean temperature (y-scale) per trip purpose (x-scale).

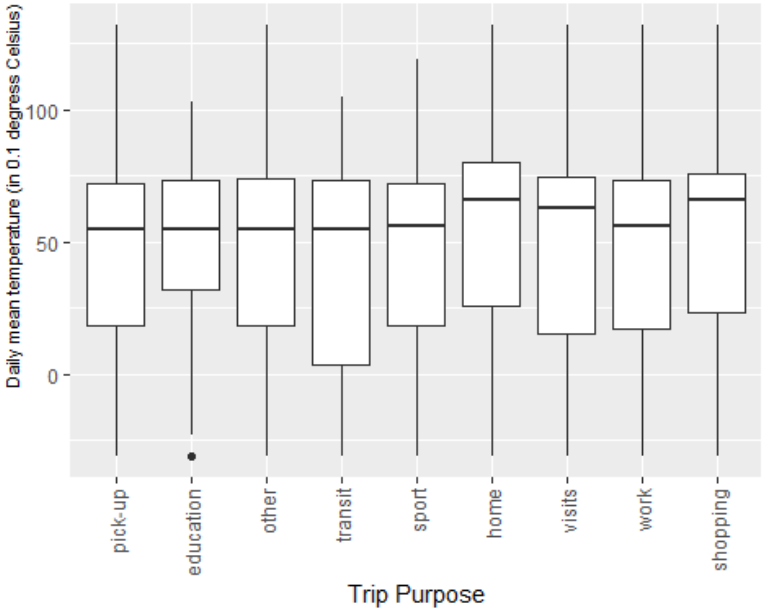
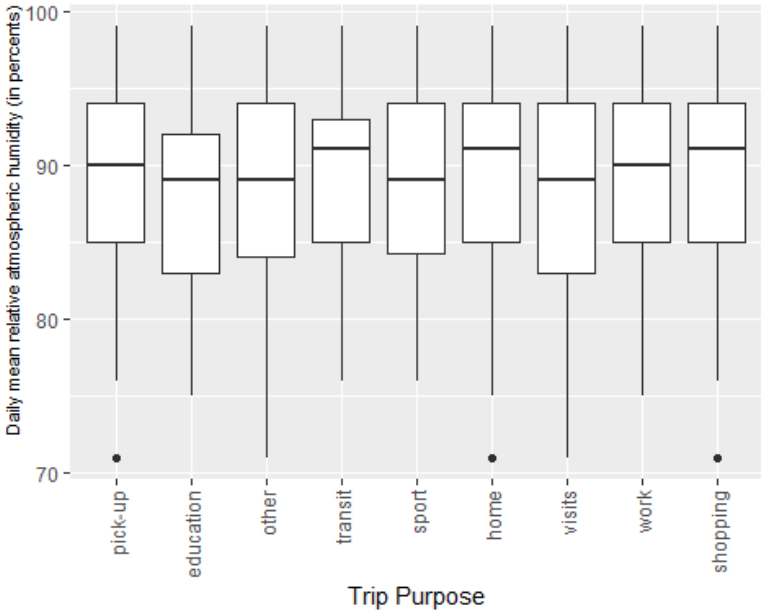


Figure A.10 Daily mean relative humidity (y-scale) per trip purpose (x-scale).



Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Grafimedia

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information