Transport Mode Detection on GPS and accelerometer data: a temporality based workflow

Thibault Fourez, Nicolas Verstaevel, Frédéric Migeon, Frédéric Schettini, Frédéric Amblard

Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2023) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2023 Montreal 17-21 July 2023

Transport Mode Detection on GPS and accelerometer data: a temporality based workflow

Thibault Fourez[a,b,*], Nicolas Verstaevel[b], Frédéric Migeon[b], Frédéric Schettini[a], Frédéric Amblard[b]

[a] *Citec Ingénieurs Conseil, 47 route des Acacias, Geneva 1711, Switzerland*

[b] *Institut de Recherche en Informatique de Toulouse, Université de Toulouse, CNRS, Toulouse INP, UT3, UT1, Toulouse, France, 118 Route de Narbonne, Toulouse CEDEX 9 31062, France*

Abstract

The knowledge of mobility in a territory is essential for local authorities' decision making. The multiplication of sensors in smartphones is an important and reliable data source to analyze users' travels. This paper presents a method for collecting GPS and accelerometer data and a processing workflow to classify users' transportation mode with high accuracy. Recommendations for collecting the measurements are explained. The pre-processing method presented is based on the analysis of time gaps between successive observations. The fusion of GPS and accelerometer data and the calculation of features are performed with a sliding time window. OCC-Transport Mode, a dataset collected with a smartphone application is presented and made public to illustrate the different steps. The performances of several classifiers are compared on the collected dataset and on two public datasets (GeoLife and US-Transportation Mode). The classification accuracy, improved by the joint use of GPS and accelerometer, is close to 100% on the collected dataset. The features resulting from the time gaps are more important than the other features in the classification. The results obtained on two public datasets are discussed.

Keywords: Transport Mode Detection; Smartphone sensors; GPS; Accelerometer

1. Introduction

The generalization and multiplication of sensors carried by human beings, via smartphones and other connected objects in particular, represent a great opportunity for local decision-makers to measure and model the impact of urban planning and public transport policies (Sauerländer-Biebl et al. (2017)). Urban planning and transportation development choices have a direct impact on the quality of life of citizens. It is therefore essential for local governments to know the state of user mobility in their territory (Bachir et al. (2019)), that is, to answer 5 main questions:

* Corresponding author.
E-mail address: thibault.fourez@irit.fr

1. **Who** are the people moving?
2. **Where** do they move?
3. **When** do they move?
4. **How** do they move?
5. **Why** do they move?

The answers to these questions are the subject of mobility surveys conducted by local authorities on large numbers of citizens. However, these surveys are infrequent (two to three times a decade) because they are costly and essentially declarative (the data come from users' declarations and are not observed nor cross-checked in an accurate and objective manner) (Bachir et al. (2019) and Gong et al. (2012)). A major challenge for local authorities is therefore to have a good estimate of the mobility as it would be given by a large-scale survey, over a given period and a given territory.

In order to ensure the objectivity of the data, recent approaches propose to use the data of sensors embedded on users (here noted "user data") to analyze their movements (Antar et al. (2021)), their driving behavior (Azadani and Boukerche (2021)) or even the state of the infrastructure they use (Yang et al. (2021)). The objective is to perform these analyses on a sufficiently large number of users to obtain an overall view representative of the real mobility. The most commonly used sensors for travel analysis are the GPS and the accelerometer embedded in all modern smartphones.

This paper focuses on the fourth question, i.e. how to implement an efficient processing workflow to detect users' transportation modes from GPS and accelerometer data. The answer to this question results in two contributions:

- A new methodology for solving the problem of supervised mode classification, based in particular on the analysis of temporal gaps between observations.
- A set of recommendations concerning the acquisition of position and acceleration data, in the form of hypotheses to be verified, to obtain a good classification of the transport mode.

These two points are illustrated using our dataset Occitania-Transportation Mode collected as part of an urban mobility research project. The methodology is then evaluated on two public datasets.

The paper in structured as follows: in section 2, related work on transport mode detection with user data is described. The characteristics and hypotheses of our dataset are presented in section 3. In the section 4 we introduce and detail a data pre-processing workflow incorporating the temporal differences of the observations. Experiments showing the efficiency of the methodology are then presented in the section 5, and their results are analyzed and discussed in the section 6.

2. Related work

This section presents different approaches to solving the mode detection problem. Two types of mobility data can be used for this purpose:

- User data which are time series of observations of a user (e.g. GPS, accelerometer).
- Contextual data which are additional information on the spatial context (GIS data), temporal context (e.g. public transport timetables, opening hours of public places) or spatio-temporal context (e.g. real time access to the position of buses or taxis).

2.1. GPS data

Visualization and analysis of user trajectories are regularly used to determine the transportation modes. The knowledge of the time series of the coordinates $(l_i)_{i\in\mathbb{N}}$ at the instants $(t_i)_{i\in\mathbb{N}}$ allows to compute attributes of speed and Bearing (direction of the user with respect to North). The distribution of the values of these attributes strongly depends on the mode of transport used. Statistical indicators (or features) of these attributes are used as input to a classifier to predict the transportation mode (Zheng et al. (2008), Dalumpines and Scott (2017) and Gonzalez et al. (2010)).

The coordinates themselves can also be used to classify the mode of transportation when put into the spatial context of the study area, i.e., when GPS data is combined with contextual data. Previous research has shown that the integration of GIS data such as road sections (Gong et al. (2012), Stenneth et al. (2011) and Biljecki et al. (2013)) or points of interest (POI) (Siła-Nowicka et al. (2016)) improves the classification of the transportation mode. Data from transit operators such as bus stops (Stenneth et al. (2011) and Siła-Nowicka et al. (2016)) or even bus positions in real time (Stenneth et al. (2011)).

Recommendations for the construction of a GPS dataset (e.g. user representativeness, acquisition) are presented in Stopher and Wargelin (2010) and lay the foundation for the creation of future datasets.

2.2. Accelerometer data

The widespread use of accelerometer in modern smartphones allows the use of the three-axis acceleration signals $(a_i)_{i\in\mathbb{N}} = ((a_i^x, a_i^y, a_i^z))_{i\in\mathbb{N}}$ at time $(t_i)_{i\in\mathbb{N}}$ to characterize the user's movements (Human Activity Recognition or HAR). The mode classification problem can be seen as an extension of HAR. Resolutions from data collected via smartphone applications (Alotaibi (2020) and Nick et al. (2010)) or from public datasets (Iskanderov and Guvensan (2020)) have been widely explored in recent years. The use of accelerometer, and specifically acceleration magnitude computation, generally provides better classification accuracy than GPS data at a lower collection energy cost.

Accelerometer data can be enriched with measurements from other motion sensors present in modern smartphones such as gyroscope or rotation vector (Jahangiri and Rakha (2014)).

2.3. Hybrid data

New approaches propose to combine accelerometer and GPS data and show improved classifier performance over the use of only one of the two sources (Widhalm et al. (2012), Reddy et al. (2010), Feng and Timmermans (2013), Shafique and Hato (2016) and Sauerländer-Biebl et al. (2017)).

Recommendations on the acquisition frequency of GPS and accelerometer data are given in Shafique and Hato (2016). The authors show that the classification accuracy and the computational cost increase jointly when the frequency of measurements decreases. It is therefore necessary to find a balance between prediction quality and the cost of associated pre-processing. The choice of the training sample size of the classifiers is studied in Shafique and Hato (2015) for the transportation mode classification problem.

The references available in the literature on the joint use of GPS and accelerometer use private datasets usually collected using private smartphone applications. The data acquisition rules are rarely detailed, making it impossible to reproduce the construction of the dataset and the associated results. Furthermore, to our knowledge, there is no attempt to list and study a set of hybrid data acquisition rules and assumptions. The methodology presented in this study needs to be made explicit to ensure the reproducibility of the results and analysis on a new territory.

In addition, several authors have attempted to exploit the frequencies of observation times $(t_i)_{i\in\mathbb{N}}$. In Biljecki et al. (2013), the time since the last stop is used as a model variable. In Feng and Timmermans (2013), the variable STEPS, which is the average time of immobility per minute, is introduced. To the best of our knowledge, there is no pre-processing workflow in the literature based on the exploitation of temporal gaps between observations as introduced in this paper.

In the next section, we introduce the collected GPS and accelerometer dataset we use to illustrate our transport mode detection methodology.

3. Collected dataset

This section presents the acquisition rules of both GPS and accelerometer signals in our dataset OCC-Transportation Mode (OCC-TMD), before detailing the hypotheses of its collection process. Raw and processed data are made publicly available in Fourez (2022).

The data was collected using a smartphone application developed as part of a mobility research project. This application passively collects GPS positions and accelerometer signals from the smartphone. This study focuses on data collected by a 25-year-old male user. This user then added a label corresponding to the mode of transportation

used for each observation point (c.f. section 3.2.2). The smartphone used for collection is a Samsung Galaxy A32 with the Android 11 operating system.

3.1. Data acquisition rules

3.1.1. GPS

Table 1. GPS data attributes

Attribute	**Description**	**Unit**
`timestamp`	Date and time	Unix timestamp (*ms*)
`lat`	Latitude	WGS 1984 coordinate system
`lon`	Longitude	WGS 1984 coordinate system
`location_accuracy`	Coordinate accuracy	*m* (meters)
`location_speed`	Speed (from previous point)	$m.s^{-1}$ (meters per second)
`location_speed_accuracy`	Speed accuracy	$m.s^{-1}$ (meters per second)
`location_heading`	Bearing (w.r.t. North)	degrees [0,360)

The attributes collected by the GPS module of the smartphone are the user's coordinates (`lat`, `lon`) and their precision (`location_accuracy`), the acquisition time (`timestamp`), the speed (`location_speed`) and its precision (`location_speed_accuracy`) and the Bearing (`location_heading`). The units of each attribute are detailed in table 1.

User location data is collected when the following two conditions are met:

- The user has moved at least 50 meters from the previous point.
- It has been at least 10 seconds since the previous point.

In other words, no two consecutive points can be less than 10 seconds apart or less than 50 meters apart. The purpose of these criteria is to save the smartphone's power by avoiding the acquisition of useless points (i.e. during moments of immobility). The time and distance thresholds were set empirically based on user feedback.

3.1.2. Accelerometer

Table 2. Accelerometer data attributes

Attribute	**Description**	**Unit**
`timestamp`	Date and time	Unix timestamp (*ms*)
`x`	Acceleration on the x-axis corrected by Gx	$m.s^{-2}$ (meter per second squared)
`y`	Acceleration on the y-axis corrected by Gy	$m.s^{-2}$ (meter per second squared)
`z`	Acceleration on the z-axis corrected by Gz	$m.s^{-2}$ (meter per second squared)

The accelerometer module of the smartphone returns the acceleration in the three axes (x, y and z) corrected for gravity, as well as the acquisition time (timestamp). The units are detailed in the table 2.

As with location, the user's accelerometer data is collected when:

- The detected acceleration is at least 1 $m.s^{-2}$ away from the previous point.
- At least 1 second has elapsed since the previous point.

A user's accelerometer signal is therefore composed of points at least one second apart. The acceleration threshold refers to the difference in acceleration between two consecutive points, not the acceleration itself. Thus, in the theoretical case where the acceleration is constant, no new points will be recorded despite the variation in the user's potential speed. As in the case of localization, these thresholds were set empirically based on user feedback.

3.2. Data collection

3.2.1. Acquisition period

The data was collected by a single user in a discontinuous manner from July 26, 2022 to August 10, 2022. This user moved around the Occitania region in the south of France (between Toulouse and Montpellier), noting for each trip the start time, end time, and mode of transportation used.

Fig. 1. Temporal distribution of points by transportation modes

Hypothesis 1 (Data continuity) *Let us consider an observation period* $\mathcal{T}_k = [t_k^0, t_k^1]$. *At any time* $t \in \mathcal{T}_k$, *the user's movements are collected, i.e. he carries a smartphone at all times with the app running.*

To satisfy the hypothesis 1, the user's movements were tracked over 4 observation periods $\mathcal{T}_1$ to $\mathcal{T}_4$ in a continuous manner. In total, the dataset contains **73 hours and 37 minutes** of travel, including 44 hours and 43 minutes for the $\mathcal{T}_1$ period, 26 hours and 6 minutes for the $\mathcal{T}_2$ period, 1 hour and 47 minutes for the $\mathcal{T}_3$ period, and 1 hour for the $\mathcal{T}_4$ period.

Table 3. Number of points by observation period and by transportation mode.

Observation phase	still	walk	bike	car	bus	metro	train	Total
$\mathcal{T}_1$	3572	6169	1037	0	1181	947	0	12906
$\mathcal{T}_2$	4892	10944	385	3789	0	0	343	20353
$\mathcal{T}_3$	694	92	0	206	0	0	877	1869
$\mathcal{T}_4$	0	858	0	0	527	239	0	1624
Total	9158	18063	1422	3995	1708	1186	1220	**36752**

3.2.2. Labeling

In the days following each observation period, the user reviewed his spatial coordinate history over the period using a GIS software. Each coordinate point was associated with one of the following 7 transportation modes: *still*, *walk*, *bike*, *car* (car), *bus*, *metro* (subway) and *train*.

Hypothesis 2 (Labeling interval) *Let be two consecutive location points l_i and l_{i+1} at times t_i and t_{i+1} and modes m_i and m_{i+1} respectively. Any observation point in the interval $(t_i, t_{i+1}]$ has for mode m_{i+1}.*

In the case where the mode changes between two consecutive location points, the label of a point corresponds to the mode used between the previous point **excluded** and itself **included** under the hypothesis 2. The labels associated with the location points are then propagated to the accelerometer points using the associated timestamp. Figure 2 shows the distribution of observation points (GPS + accelerometer) by transport mode. The numbers of points collected by observation period and label are presented in table 3.

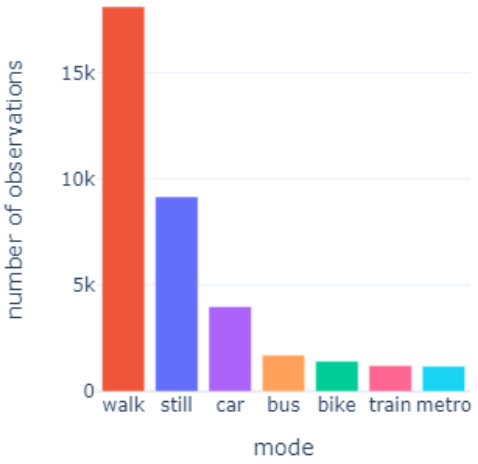

Fig. 2. Bar chart of the number of points by transportation mode.

The acquisition rules and hypotheses presented in this section are a contribution to the implementation of a transportation mode detection system based on a smartphone data collection program. The next section details the pre-processing of the GPS and accelerometer data collected on users. The different steps are illustrated using the collected dataset.

4. Pre-processing

This section presents the data pre-processing chain to extract statistical features from the raw data collected by the user. The raw data contains 40830 accelerometer points and 1168 GPS points across the 4 observation periods, for a total of 41998 user observation points.

4.1. Data filtering

The location data is provided with an accuracy field (`locationAccuracy`). In order to eliminate outlier GPS observations, points whose location accuracy is strictly greater than 100 meters are arbitrarily removed. The number of accelerometer points remains unchanged, but 18 location points are eliminated. After the data filtering step, a total of **41980 observation points** remain.

4.2. Adding labels

The data are collected over observation periods $(\mathcal{T}_k)_{k\in[\![1,4]\!]}$. The labels were defined by the user from the location points. An observation period $\mathcal{T}_k = [t_k^0, t_k^1]$ is divided by the user into segments $(t_i, t_{i+1}]$ characterized by a starting

localization point l_i, an ending localization point l_{i+1}, and a mode m_{i+1} (c.f. hypothesis 2). Each location point is thus labeled with the mode of the segment to which it belongs, except for **14 points** which are not included in any segment (they are outside any observation period). On the other hand, the accelerometer points a_i have a different (and much higher) acquisition frequency than the GPS points. They are therefore not necessarily included in the observation periods, and thus in the segments $(t_i, t_{i+1}]$ labelled by the user. The table 4 illustrates how acceleration points not belonging to any segment are removed from the dataset. **5153 accelerometer points** are not detected in any segment. At the end of this step, **36813 observation points** remain.

Table 4. Example of label propagation to the accelerometer points. Only the points a_3, a_4,a_7, a_8, a_9 and a_{10} are included in an observation period and are associated with a label. The other acceleration points have no known mode and are removed from the dataset.

a_1	a_2	l_1^0	a_3	a_4	l_1^1	a_5	l_2^0	l_2^1	a_6	l_3^0	a_7	a_8	a_9	l_i	l_3^1	l_4^0	a_{10}	l_4^1	a_{11}
?	?	$\mathcal{T}_1$				?	$\mathcal{T}_2$		?	$\mathcal{T}_3$						$\mathcal{T}_4$			?

4.3. Attribute calculation

In addition to the attributes of the location and accelerometer points, 4 additional attributes (time gap and distance for the location points, time gap and magnitude for the accelerometer points) are calculated. The magnitude of the accelerometer is calculated from the following formula:

$$\texttt{accelerometer\_magnitude} = \sqrt{\texttt{x}^2 + \texttt{y}^2 + \texttt{z}^2}$$

where `x`, `y` and `z` are the corrected acceleration values in the three axes (c.f. table 2).

Descriptions and units of the additional attributes are given in table 5.

Table 5. Additional attributes computed on the location points (first 2 rows) and on the accelerometer points (last 2 rows).

Attribute	Description	Unit
`location_time_delta`	Time gap from previous point	s (seconds)
`location_distance_delta`	Distance from previous point	m (meters)
`accelerometer_time_delta`	Time gap from previous point	s (seconds)
`accelerometer_magnitude`	Accelerometer magnitude	$m.s^{-2}$

The attributes of time gap or distance require the calculations to be performed separately for each observation period to respect the hypothesis 1 of data continuity. A time or distance value between the last point of one period and the first of the next period is meaningless because the user has potentially made many other trips for which there are no observations. Thus, for each observation period, the first location point and the first accelerometer point are removed from the dataset because no differential attribute can be computed with respect to the previous point (i.e., **8 points removed** in total). At the end of the calculation of the additional attributes, the dataset contains **36805 observation points**.

4.4. Data fusion and feature computation

At this point, the dataset contains location points and accelerometer points with different attributes. In order to train classifiers, it is necessary to merge these two data sources to compute the same statistical features on all the points. For each observation period, the data fusion step and the feature computation step are performed simultaneously through a sliding window according to the following process:

1. Concatenation of n_k^l location points $(l_i)_{i\in[\![1,n_k^l]\!]}$ and n_k^a acceleration points $(a_i)_{i\in[\![1,n_k^a]\!]}$ of the observation period $\mathcal{T}_k$ into a single ordered time series of n_k observation points at times $(t_i)_{i\in[\![1,n_k]\!]}$.
2. Slide of the whole data with time windows $T_i = (t_j)_{j\in[\![1,n_k]\!],|t_i-t_j|\leq\frac{T}{2}}$ of duration T centered around the observation points at time t_i.
3. For each observation point at time t_i, computation of the p statistical features $(f_j(T_i))_{j\in[\![1,p]\!]}$ over the time window T_i.
4. Completion of missing data by backward fill (i.e. value of the next point).

The purpose of using time windows in step 2 is to gather the observation points in the same temporal neighborhood to obtain location and accelerometer information. Statistical features are then computed on this temporal neighborhood. The duration T of the temporal neighborhood must be large enough to capture the variation of the acceleration (whose minimum frequency is 1 second) and the localization information, while being shorter than the duration of a typical move to avoid confusion between modes in the statistical computation of features. A duration of 60 seconds seems to meet these criteria. With such a duration, 14 observation points have no accelerometer information in their time window (i.e., no statistical features from the accelerometer attributes), and 12367 observation points have no location information, which is 33.6

The features computed in step 3 are the 5 statistical indicators `min` (minimum), `max` (maximum), `mean` (average), `median` (median) and `std` (standard deviation) computed on 4 basic location attributes (`speed`, `speedAccuracy`, `locationAccuracy`, and `heading` (bearing)) and the 4 additional attributes described in section 4.3. In total, **40 statistical features** are computed for each observation point.

In order to handle missing data in step 4, a backward fill is performed, i.e. for each missing attribute the first value found in the following points is propagated. The choice of the filling direction comes from the hypothesis 2 which implies that the mode at an unobserved time t is the mode of the point observed at time $t_i = \min\{t_i \mid i \in [\![1,n]\!], t < t_i\}$, i.e. the first point observed after time t.

At the end of this process, the observation points of each period are gathered. By construction, each observation period has as its last element a location point (c.f. section 4.2). In the case where this point has no accelerometer points in its temporal neighborhood (and thus missing values for features derived from accelerometer information), it is removed from the dataset. In the user-collected data, this was not the case for any of the 4 observation periods. Therefore, the final dataset is composed of $n = 36805$ **rows (observation points), and** $p = 40$ **columns (statistical features)** plus the transport mode and the observation period.

The pre-processing step presented in this section integrates a temporal dimension to the construction of the training variables of a classification model. On the one hand, a time gap attribute between successive points is computed for each data signal, and on the other hand the data fusion and the computation of statistical features on the set of attributes is performed using a sliding time window. Both raw and processed data from OCC-TMD are made public in Fourez (2022). The following section presents the classification of the transport mode based on the features computed on the collected dataset and on two other datasets from the literature.

5. Transport mode classification

Transportation mode detection can be viewed as a supervised classification problem. The individuals to be classified are the observation points at the output of the pre-processing described in section 4, i.e. vectors of dimension $p = 40$, i.e. the number of statistical features computed from the location and accelerometer information.

In the literature, there is a large variety of classifiers used to solve the mode detection problem. A first approach consists in defining an algorithm based on rules from the mobility analysis domain (Siła-Nowicka et al. (2016)). In Gong et al. (2012), these rules are based on measurements and on the integration of GIS data. In Biljecki et al. (2013), the authors use a fuzzy expert system as classifier.

Many references use Machine Learning techniques. The goal is to search by learning on data $(X_{\text{train}}, y_{\text{train}})$ an estimate $\hat{f}$ of the theoretical objective function f such that for any observation $x_i \in X_{\text{train}}$ and its associated mode of

transport $y_i \in y_{\text{train}}$, the following relation is verified:

$$y_i = f(x_i)$$

Once the function $\hat{f}$ is learned by the Machine Learning algorithm, an estimate $\hat{y} = \hat{f}(X)$ of the transport modes y of new observations X can be obtained. Among the most used algorithms in this framework, we can list Logistic Regression (Logit) (Dalumpines and Scott (2017)), Decision Trees (DT) (Stenneth et al. (2011), Reddy et al. (2010), Zheng et al. (2008) and Alotaibi (2020)), Naive Bayesian classifiers (NB) (Stenneth et al. (2011) and Nick et al. (2010)), Bayesian Networks (BN) (Stenneth et al. (2011) and Zheng et al. (2008)), Support Vector Machines (SVM) (Jahangiri and Rakha (2014) and Nick et al. (2010)) and Hidden Markov Models (HMM) (Widhalm et al. (2012) and Reddy et al. (2010)). Ensemble approaches are also studied, including bagging (Alotaibi (2020), Random Forests (RF) (Stenneth et al. (2011), Alotaibi (2020) and Shafique and Hato (2016)) and the Gradient Boosting algorithm (Alotaibi (2020)). Dense Artificial Neural Networks (ANN) (Stenneth et al. (2011) and Gonzalez et al. (2010)) and Recurrent Neural Networks (LSTM) (Iskanderov and Guvensan (2020)) are regularly used in the literature.

We propose to compare the performances of several Machine Learning algorithms from the literature. In the rest of this section, the experiment is replicated on two public datasets: Microsoft GeoLife for GPS data, and US-TMD for accelerometer data.

5.1. OCC-Transportation Mode dataset

OCC-TMD combines location and accelerometer information. In order to study the contribution of each type of information to the detection of the transport mode, it is necessary to define several subsets of the dataset for the experiments. The complete dataset is denoted D_{full}. The datasets with features from the location and accelerometer data are denoted D_{location} and $D_{\text{accelerometer}}$ respectively. Moreover, in order to compare the results with those from the GeoLife dataset (c.f. section 5.2), the D'_{location} dataset containing the features from the localization without the position or speed accuracy information is defined. The input variables of the mode classification problem for each dataset are presented in the table 6.

Table 6. Subsets of features of the different variants of the dataset. 5 features are computed from each attribute (mean, minimum, maximum, standard deviation and median).

Attribute	D_{full}	D_{location}	D'_{location}	$D_{\text{accelerometer}}$
`location_accuracy`	✓	✓		
`location_speed`	✓	✓	✓	
`location_speed_accuracy`	✓	✓		
`location_heading`	✓	✓	✓	
`location_distance_delta`	✓	✓	✓	
`location_time_delta`	✓	✓	✓	
`accelerometer_magnitude`	✓			✓
`accelerometer_time_delta`	✓			✓
Number of features	**40**	**30**	**20**	**10**

On each subset, the performances of 4 classifiers from Machine Learning are compared for solving the transportation mode classification problem:

- KNN : k Nearest Neighbors (Cover and Hart (1967))
- RF : Random Forest (Breiman (2001))
- ANN : Artificial Neural Network (Rumelhart et al. (1986))
- SVM : Support Vector Machine (Cristianini et al. (2000))

The used implementations come from the python library *scikit-learn* (Pedregosa et al. (2011)). For each subset, the parameters of each classifier are chosen by 5-fold cross-validation from the parameter grids presented in the table 7.

Table 7. Grids of values for parameter optimization by cross-validation. The other parameters keep their default values in the implementation of *scikit-learn*.

Parameter	**Grid of values**		
	KNN		
`n_neighbors`	5	10	15
`weights`	uniform	distance	
	RF		
`n_estimators`	100	200	300
	ANN		
`hidden_layer_sizes`	100	200	500
`alpha`	1e-4	1e-3	1e-2
`learning_rate_init`	1e-3	1e-2	
	SVM		
`kernel`	rbf	poly	
`C`	1	50	100

5.2. *GeoLife dataset*

GeoLife (Zheng et al. (2009)) is a dataset created by Microsoft Research Asia between April 2007 and August 2012. It compiles the GPS trajectories of nearly 180 users over a total of more than 48,000 hours, collected with GPS trackers or GPS-enabled phones. Each point of the dataset is defined by GPS coordinates (latitude, longitude, altitude) and a transportation mode (walking, cycling, bus, car, metro, train, plane, boat, race, motorcycle or taxi).

In order to compare GeoLife with the collected dataset, we have to apply the pre-processing presented in section 4:

1. Filtering of the unlabeled data or with modes absent from the collected dataset (plane, boat, race, motorcycle and cab).
2. Computation of the additional location attributes.
3. Computation of statistical features on the attributes (mean, minimum, maximum, standard deviation and median) with sliding windows of duration $T = 60$ seconds.

After filtering the data in step 1, the dataset contains 5.17 million observation points. In contrast to the collected dataset, the *still* mode (no travel) is absent.

Besides the additional attributes `location_time_delta` and `location_distance_delta`, an estimation of the attributes `location_speed` and `location_heading` between two consecutive points is computed in step 2 because the GeoLife dataset does not contain any speed or bearing information. The calculation of the attributes must be done for each observation period. The GeoLife data is segmented by observation periods at each user change or at each time difference of more than 24 hours. 890 observation periods are obtained, with a mean number of 5809 observation points and a median number of 1504 observation points. After the calculation of the attributes, the observation points whose estimated speed is higher than 400km/h are filtered out.

The features from the location attributes are computed in the same way as for the collected dataset in step 3. The dataset (noted D_{geolife}) contains 4.7 million observation points at the end of the pre-processing. The two most frequently observed modes are *walk* and *bus*. The distribution of the other modes is shown in Figure 3.

Due to a much larger number of points than in the collected dataset, a $D^{\text{test}}_{\text{geolife}}$ subset of 40,000 points is randomly selected for classifier training. Among these observation points, 80% are used for training and 20% for testing the models. The classification accuracy of each model introduced in section 5.1 is then evaluated on the whole dataset. For each algorithm, the combination of parameters selected by cross-validation is used. Moreover, in order to evaluate

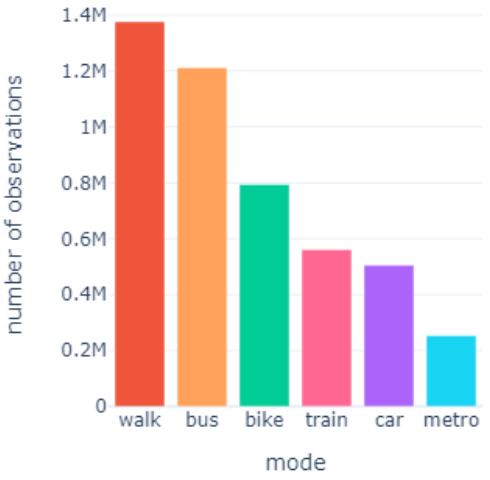

Fig. 3. Bar chart of the number of points by transportation mode in the GeoLife dataset after pre-processing.

the contribution of the differential attributes (`location_time_delta` and `location_distance_delta`), the subsets D'_{geolife} and D'_{geolife} deprived of the features computed from these two attributes are defined.

5.3. US-Transportation Mode dataset

US-Transportation Mode (US-TMD) (Carpineti et al. (2018)) is a dataset collected by the University of Bologna with a smartphone tracking application used by 13 volunteers. Each user launched several acquisitions corresponding to some of their trips. The collected points are then labeled with the transport mode used during that period (still, walk, car, bus or train).

From the raw data, the same pre-processing as for the collected dataset is applied:

1. Filtering of the data by validity and by sensor type.
2. Calculation of the accelerometer attributes (magnitude and temporal gaps) on each observation period.
3. Computation of statistical features on the attributes (mean, minimum, maximum, standard deviation and median) with sliding windows of duration $T = 60$ seconds for each observation period.

The raw data contains measurements acquired by different sensors of the users' smartphones. In this experiment, only the accelerometer and speed sensors are retained in order to test the quality of the implemented processing chain (GPS positions are not provided). However, due to the very low number of available speed values, only the information from the accelerometer is kept. After selecting the observation points from the accelerometer and deleting the missing or invalid data, the dataset contains 1.4 million points.

The observation periods are defined in the same way as for the GeoLife dataset (see section 5.2), with the addition of a segmentation between the different acquisition files of the same user. Thus, each observation period has a unique transport mode.

At the end of the feature computation, the dataset (noted D_{ustmd}) contains 1.4 million points. The distribution of the transport modes is illustrated in figure 4. Unlike the collected dataset, the metro mode is absent.

As with the GeoLife dataset, a subset $D^{\text{test}}_{\text{ustmd}}$ of 40,000 points (of which 80% for learning and 20% for testing) is randomly selected to train the models before studying their performance on the full dataset. We also define the subsets

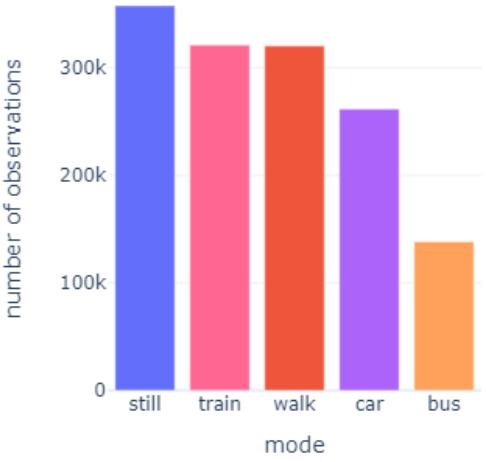

Fig. 4. Bar chart of the number of points by transportation mode in the US-TMD dataset after pre-processing.

D'_{ustmd} and $D'^{\text{test}}_{\text{ustmd}}$ deprived of the features issued from the attribute `accelerometer_time_delta` in order to study their relevance.

6. Results

This section presents the results obtained from transport mode classification with the classifiers presented in section 5 for each of the three datasets. A quantitative and qualitative analysis of the relevance of using time gaps is also performed.

6.1. Collected dataset

Table 8. Comparison of classification accuracies. For the first 4 classifiers, the best accuracy on the cross-validation presented in section 5.1 was retained.

	D_{full}	D_{location}	D'_{location}	$D_{\text{accelerometer}}$
RF	**99.57%**	**97.06%**	**97.02%**	**98.65%**
SVM	96.57%	94.36%	88.91%	74.49%
ANN	98.14%	95.24%	90.41%	85.33%
KNN	99.01%	**96.44%**	96.41%	**94.79%**

Table 8 presents the classification accuracies obtained for the collected dataset according to the following formula:

$$\text{accuracy} = \frac{\text{number of correct predictions}}{\text{number of predictions}}$$

The Random Forest classifier systematically obtains the best accuracy, whatever the subset. When all features are used, the algorithm discriminates the transport mode with 99.57% accuracy. The use of SVM gives the lowest performance, especially when only the features from the accelerometer data are considered (74.49% accuracy).

For each classifier, the best performance is achieved with the full D_{full} subset. Accelerometer data seem to be less discriminating than GPS data because the accuracy obtained with the $D_{\text{accelerometer}}$ dataset is lower than that obtained with D_{location} (except with Random Forest). Among GPS data, the inclusion of precision attributes in D_{location} systematically increases the precision obtained with D'_{location}. For all classifiers, the best accuracy is obtained by using GPS and accelerometer data together.

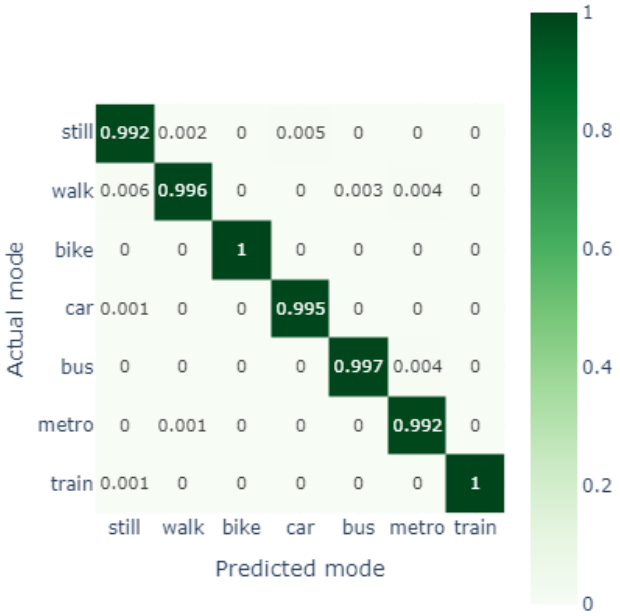

Fig. 5. Normalized confusion matrix with the Random Forest classifier on the D_{full} dataset.

The confusion matrix presented in Figure 5 shows very small errors for each predicted mode. The two modes with the most confusion (0.6%) are Still and Metro. These two modes are characterized by a low number of observations, by acquisition rule for the Still mode and by signal loss for the Metro mode.

The importance of the features is evaluated with the Gini impurity measure computed during the training of the Random Forest instance that gave the best accuracy. The objective is to evaluate the relevance of the pre-processing chain, and in particular the computation of the time gap attributes. The importance values are illustrated in the figure 6. The most important feature is `std_location_distance_delta`, that is, the standard deviation of the distance between two successive GPS points. Of all the features, those resulting from the time gaps have a higher average importance than the others (0.029577 versus 0.023474). The features coming from the accelerometer have a slightly higher average importance than the others (0.025788 against 0.024737), which explains the higher performance of the Random Forest classifier on the subset $D_{textaccelerometer}$.

The next two features in order of importance are derived from the time gap attributes in the accelerometer and GPS data. In order to visualize the discriminating power of these two features, their distributions for each mode are presented in figure 7 as boxplots. It appears that the time gaps in the accelerometer data easily discriminate between Train and Bus modes, while those in the GPS data significantly separate Train, Metro and Still modes from the other modes. This is due to the difference in observation frequency for these modes in particular, because of the data acquisition rules for Still (very low frequency) and Bus (very high frequency) modes, and the loss of GPS signal for Metro and Train modes.

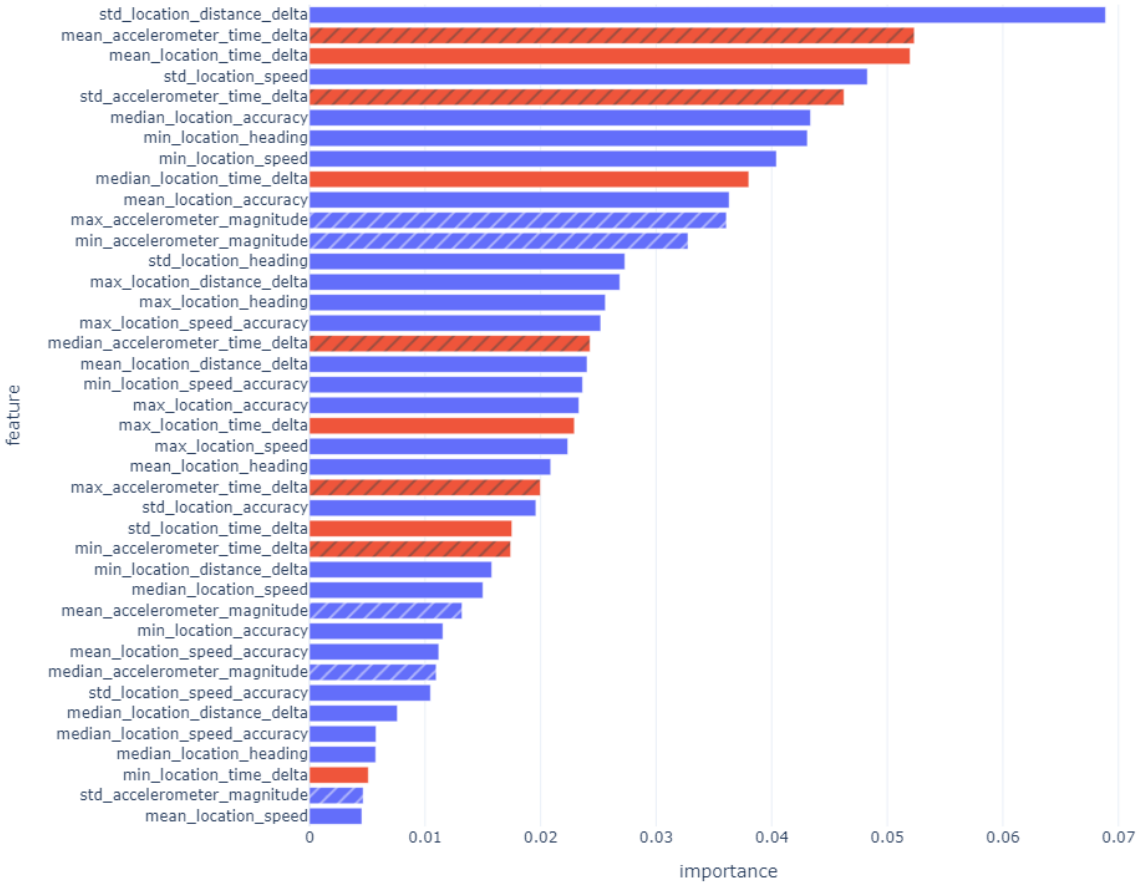

Fig. 6. Gini importances for the Random Forest classifier trained with the best combination of parameters (selected by cross-validation). Time gap features are shown in blue. Features from the accelerometer data are shown hatched.

Table 9. Comparison of classification accuracies on the different subsets of the GeoLife dataset. The accuracies of the full datasets D_{geolife} and D'_{geolife} are obtained with the models trained on the subsets $D^{\text{test}}_{\text{geolife}}$ and $D'^{\text{test}}_{\text{geolife}}$ respectively.

	$D^{\text{test}}_{\text{geolife}}$	D_{geolife}	$D'^{\text{test}}_{\text{geolife}}$	D'_{geolife}
RF	**77.39%**	70.42%	**73.08%**	**71.66%**
SVM	71.49%	69.87%	71.19%	69.81%
ANN	73.43%	**71.83%**	71.23%	70.14%
KNN	70.15%	68.87%	69.74%	68.78%

6.2. *GeoLife dataset*

The accuracies obtained with the GeoLife dataset and presented in the table 9 are maximal with the Random Forest classifier in the majority of cases (up to 77.39% with the full test dataset $D^{\text{test}}_{textgeolife}$) and minimal with the KNN classifier (70.15% with the same data) The performances on the test sub-samples are slightly higher than on the full datasets. The use of the features from the temporal gaps in the subsets $D^{\text{test}}_{\text{geolife}}$ and D_{geolife} gives systematically better performances than in the subsets $D'^{\text{test}}_{\text{geolife}}$ and D'_{geolife} that do not include them.

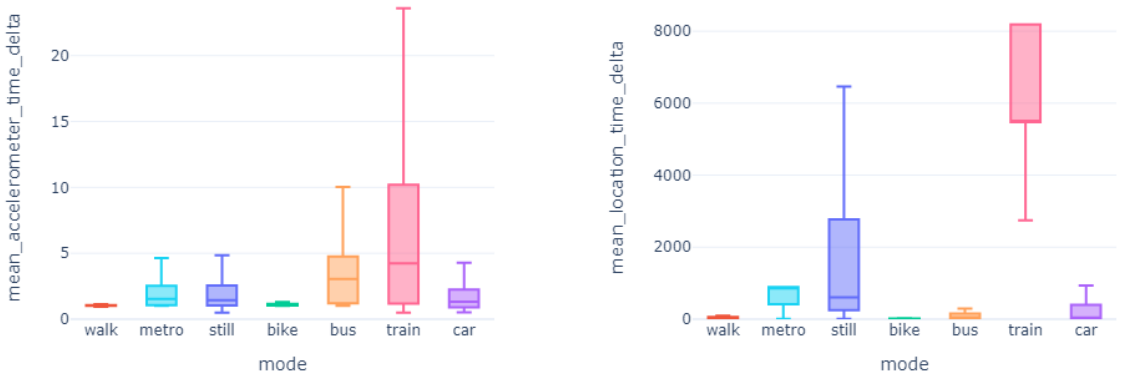

Fig. 7. Distribution of features `mean_accelerometer_time_delta` (left) and `mean_location_time_delta` (right) by transportation mode. The data is represented as boxplots without the outliers.

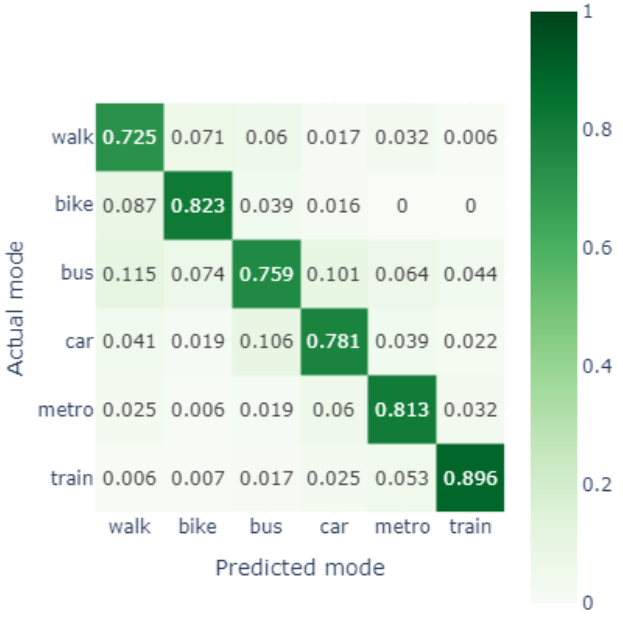

Fig. 8. Normalized confusion matrix with the Random Forest classifier on the $D^{\text{test}}_{\text{geolife}}$ dataset.

The confusion matrix obtained with the Random Forest classifier (c.f. figure 8) allows to evaluate the prediction errors committed for each predicted class. These errors are almost always lower than 10%, except in the case of the Walk mode (11.5% of the predictions correspond to the Bus mode) and the Car mode (10.1% of the predictions correspond to the Bus mode). The Car/Bus confusion can be explained by similar traffic speeds, while the Car and Walk mode trips are characterized by still moments (the GeoLife dataset does not have a Still class).

6.3. US-Transportation Mode dataset

The results obtained with the US-TMD dataset confirm those obtained with the GeoLife dataset. As illustrated in the table 10, Random Forest is among the best performing classifiers (alongside KNN), with a maximum accuracy of

Table 10. Comparison of classification accuracies on the different subsets of the US-TMD dataset with the accuracies obtained in Carpineti et al. (2018) and Vakili et al. (2020). The accuracies of the full datasets D_{ustmd} and D'_{ustmd} are obtained with the models trained on the subsets $D^{\text{test}}_{\text{ustmd}}$ and $D'^{\text{test}}_{\text{ustmd}}$ respectively.

	$D^{\text{test}}_{\text{ustmd}}$	D_{ustmd}	$D'^{\text{test}}_{\text{ustmd}}$	D'_{ustmd}	Carpineti et al. (2018)	Vakili et al. (2020)
RF	**99,39%**	96,82%	**98,46%**	96,88%	**89,00%**	**85,00%**
SVM	91,70%	91,25%	80,89%	81,56%	86,00%	79,00%
ANN	95,74%	95,04%	91,76%	91,81%	87,00%	75,00%
KNN	98,64%	**98,61%**	98,15%	**97,92%**		80,00%

99.39% achieved on the whole test dataset. As for the collected dataset, SVM gives the worst performance compared to the other classifiers (91.70% on the whole dataset). The time gap features (in the subsets $D^{texttest}_{textustmd}$ and $D_{textustmd}$) improve the accuracies obtained in almost all cases.

The results are compared to those obtained in Carpineti et al. (2018) (dataset authors) and Vakili et al. (2020). The classification accuracy is strongly improved with the previously presented approach (around 10% with Random Forest), even without using the time gaps. The parameters used in our models are those selected by cross-validation on the collected dataset. The parameters used in the models of the two articles have been optimized by their respective authors.

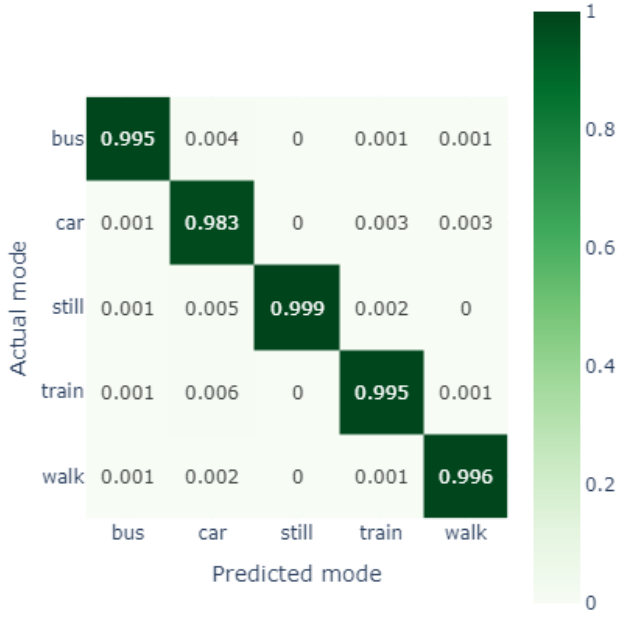

Fig. 9. Normalized confusion matrix with the Random Forest classifier on the $D^{\text{test}}_{\text{ustmd}}$ dataset.

Figure 9 presents the errors made for each predicted class. They are all lower than 1%. The accuracy obtained for the prediction of the Car mode is slightly lower than for the other modes (98.3% against more than 99.5% for all the other classes).

The results presented in this section show that the acquisition rules presented in section 3, as well as the pre-processing step based on the time gap analysis presented in section 4 are relevant for the transport mode detection. The Random Forest classifier seems to perform better in most cases. In addition, the combination of GPS and accelerometer data improves the classification accuracy compared to using only one of these two data sources.

7. Conclusion

This paper presents a new methodology for supervised transportation mode classification based on GPS and accelerometer data. It consists of the following components:

- A set of acquisition rules presented in section 3.1.
- A set of hypotheses to be respected on data collection, presented in the section 3.2.
- A data pre-processing chain explained in the section 4, based in particular on the analysis of time gaps between two successive points.

Our methodology is illustrated with the OCC-Transportation Mode dataset collected in the framework of a research project on mobility, and with the public datasets GeoLife and US-TMD. The different steps of the contributions are detailed in order to ensure their reproducibility on other territories. We also made OCC-TMD dataset public to serve as a training dataset to solve mode detection problem in future studies.

We obtain high classification accuracies on our dataset and on US-TMD which tend to validate the relevance of the presented methodology. The results show that the acquisition rules implemented as well as the use of time gaps improve the classification accuracy. However, the time gaps between the points are only exploitable when the hypothesis 1 on the continuity of the data is verified. Indeed, if the data acquisition is interrupted over a time interval, the points preceding and following this interval will have a time gap similar to the ones observed for modes such as train or subway. Moreover, the statistical distribution of time gaps highly depends on the collection process of the application and the smartphone itself, making its use difficult to generalize to other data sources. Future work should address this problem with an enhanced dataset containing data from different smartphone and software manufacturers.

In the case of the US-TMD dataset, the use of time gaps is not sufficient to explain the significant performance gain over the authors' results. Further study of the influence of the labeling method, the segmentation of trips, or the calculation of statistical features from the attributes is needed. The addition of new modes of transportation such as boat, plane or electric bikes and scooters could make the acquisition rules and the features used in the processing chain evolve. Finally, it would be interesting to deepen the notion of explainability in the prediction of the transport mode. Regarding the acquisition rules introduced, a study of the impact of the sensor collection frequency on the accuracy and energy consumption of the smartphone would allow to establish a balance between data quality and acceptability by users.

Acknowledgements

We thank the National Association for Research and Technology (ANRT) for the CIFRE funding of the thesis project in partnership with the Institut de Recherche en Informatique de Toulouse (IRIT) and the company Citec Ingénieurs Conseil. We also thank all the reviewers for their help and advice.

References

Alotaibi, B., 2020. Transportation mode detection by embedded sensors based on ensemble learning. IEEE Access 8, 145552–145563.

Antar, A.D., Ahmed, M., Ahad, M.A.R., 2021. Recognition of human locomotion on various transportations fusing smartphone sensors. Pattern Recognition Letters 148, 146–153.

Azadani, M.N., Boukerche, A., 2021. Driving behavior analysis guidelines for intelligent transportation systems. IEEE Transactions on Intelligent Transportation Systems .

Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, M., Puchinger, J., 2019. Inferring dynamic origin-destination flows by transport mode using mobile phone data. Transportation Research Part C: Emerging Technologies 101, 254–275.

Biljecki, F., Ledoux, H., Van Oosterom, P., 2013. Transportation mode-based segmentation and classification of movement trajectories. International Journal of Geographical Information Science 27, 385–407.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Carpineti, C., Lomonaco, V., Bedogni, L., Di Felice, M., Bononi, L., 2018. Custom dual transportation mode detection by smartphone devices exploiting sensor diversity, in: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), IEEE. pp. 367–372.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE transactions on information theory 13, 21–27.

Cristianini, N., Shawe-Taylor, J., et al., 2000. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.
Dalumpines, R., Scott, D.M., 2017. Making mode detection transferable: extracting activity and travel episodes from gps data using the multinomial logit model and python. Transportation planning and technology 40, 523–539.
Feng, T., Timmermans, H.J., 2013. Transportation mode recognition using gps and accelerometer data. Transportation Research Part C: Emerging Technologies 37, 118–130.
Fourez, T., 2022. Occitania-transportation mode. URL: https://doi.org/10.5281/zenodo.7386788, doi:10.5281/zenodo.7386788.
Gong, H., Chen, C., Bialostozky, E., Lawson, C.T., 2012. A gps/gis method for travel mode detection in new york city. Computers, Environment and Urban Systems 36, 131–139.
Gonzalez, P.A., Weinstein, J.S., Barbeau, S.J., Labrador, M.A., Winters, P.L., Georggi, N.L., Perez, R., 2010. Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks. IET Intelligent Transport Systems 4, 37–49.
Iskanderov, J., Guvensan, M.A., 2020. Breaking the limits of transportation mode detection: Applying deep learning approach with knowledge-based features. IEEE Sensors Journal 20, 12871–12884.
Jahangiri, A., Rakha, H., 2014. Developing a support vector machine (svm) classifier for transportation mode identification by using mobile phone sensor data, in: Transportation Research Board 93rd Annual Meeting, p. 1442.
Nick, T., Coersmeier, E., Geldmacher, J., Goetze, J., 2010. Classifying means of transportation using mobile sensor data, in: The 2010 International Joint Conference on Neural Networks (IJCNN), IEEE. pp. 1–6.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.
Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., Srivastava, M., 2010. Using mobile phones to determine transportation modes. ACM Transactions on Sensor Networks (TOSN) 6, 1–27.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. nature 323, 533–536.
Sauerländer-Biebl, A., Brockfeld, E., Suske, D., Melde, E., 2017. Evaluation of a transport mode detection using fuzzy rules. Transportation research procedia 25, 591–602.
Shafique, M.A., Hato, E., 2015. Formation of training and testing datasets, for transportation mode identification. Journal of Traffic and Logistics Engineering Vol 3.
Shafique, M.A., Hato, E., 2016. Travel mode detection with varying smartphone data collection frequencies. Sensors 16, 716.
Siła-Nowicka, K., Vandrol, J., Oshan, T., Long, J.A., Demšar, U., Fotheringham, A.S., 2016. Analysis of human mobility patterns from gps trajectories and contextual information. International Journal of Geographical Information Science 30, 881–906.
Stenneth, L., Wolfson, O., Yu, P.S., Xu, B., 2011. Transportation mode detection using mobile phones and gis information, in: Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems, pp. 54–63.
Stopher, P.R., Wargelin, L., 2010. Conducting a household travel survey with gps: reports on a pilot study, in: World Congress on Transport Research, 12th, 2010, Lisbon, Portugal.
Vakili, M., Ghamsari, M., Rezaei, M., 2020. Performance analysis and comparison of machine and deep learning algorithms for iot data classification. arXiv preprint arXiv:2001.09636 .
Widhalm, P., Nitsche, P., Brändie, N., 2012. Transport mode detection with realistic smartphone sensor data, in: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), IEEE. pp. 573–576.
Yang, C.L., Sutrisno, H., Chan, A.S., Tampubolon, H., Wibowo, B.S., 2021. Identification and analysis of weather-sensitive roads based on smartphone sensor data: A case study in jakarta. Sensors 21, 2405.
Zheng, Y., Liu, L., Wang, L., Xie, X., 2008. Learning transportation mode from raw gps data for geographic applications on the web, in: Proceedings of the 17th international conference on World Wide Web, pp. 247–256.
Zheng, Y., Zhang, L., Xie, X., Ma, W.Y., 2009. Mining interesting locations and travel sequences from gps trajectories, in: Proceedings of the 18th international conference on World wide web, pp. 791–800.