European Transport Research Review

ORIGINAL PAPER Open Access

Facing the needs for clean bicycle data – a bicycle-specific approach of GPS data processing

Sven Lißner* and Stefan Huber

Abstract

Background: GPS-based cycling data are increasingly available for traffic planning these days. However, the recorded data often contain more information than simply bicycle trips. GPS tracks resulting from tracking while using other modes of transport than bike or long periods at working locations while people are still tracking are only some examples. Thus, collected bicycle GPS data need to be processed adequately to use them for transportation planning.

Results: The article presents a multi-level approach towards bicycle-specific data processing. The data processing model contains different steps of processing (data filtering, smoothing, trip segmentation, transport mode recognition, driving mode detection) to finally obtain a correct data set that contains bicycle trips, only. The validation reveals a sound accuracy of the model at its' current state (82–88%).

Keywords: Bicycle traffic planning; GPS data, Big data, Crowdsourcing, Data processing

1 Introduction

Area-wide cycling data are still hardly available and rarely used for bicycle specific traffic planning these days. Tracking cyclists routes using smartphone applications can help to fill this data gap. A big amount of data can be collected within a very short period using crowdsourcing approaches that cover hundreds or thousands of cyclists using their smartphones to track their rides. This type of data collection is not new to scientists. First approaches were made in 2007 using hand-held GPS devices (e.g. [1, 15, 23]). More expanded studies emerged with the development of more cheap GPS sensors and their integration in smartphones and their increased distribution. The studies of Charlton et al. [8], Broach et al. [5] and Jestico et al. [17] are some example in this context.

However, the collected data can contain many more information than simply bicycle trips as desired by scientists or traffic planners. Therefore, three major issues need to be considered when using smartphone-based crowd-sourced GPS data. First, the recorded data may include 'activities' at a location (e.g. paper work in the office) when cyclists or study participants forget to stop tracking after their trip already ended. Second, the recorded tracks may contain trips of other modes of transport (e.g. when people change the mode of transport and keep on tracking). At least, the recorded data often contains so-called 'noisy data', which occur because of the functionality of the GPS system itself (e.g. loss of signals or diffraction of signals, which leads to GPS point jumping). All these issues occur when cyclist record their trips using GPS and smartphone applications.

Aim of the study was to develop a bicycle specific data processing approach, which is capable to process big GPS data sets and easy to use and to implement

* Correspondence: sven.lissner@tu-dresden.de
Technische Universität Dresden, Chair of Transportation Ecology, Hettnerstraße 1, 01062 Dresden, Germany

for practitioners. The goal was to create a method which is highly transparent, flexible and interpretable (no black box). Furthermore a high accuracy is an essential requirement. Therefore, the article presents an approach of comprehensive GPS data processing. Existing work is briefly highlighted and discussed in the following section (2). Section 3 contains methodology and the description of the data that has been used to develop and to validate the developed data processing approach. We present our main findings in section 4 and close the paper by discussing the results and drawing further research opportunities in sections 5 and 6.

2 State of research

Scientists and practitioners widely acknowledge the need for data pre-processing of GPS-based and crowdsourced traffic data before further using the data, for instance to estimate route choice models. In the most common studies, data-pre-processing consists of three main steps.

So-called noisy data (e.g. containing GPS outliers) is reduced in a first step. This is mostly done with very basic threshold filters using speed parameters to reduce GPS leaps (see for example the studies of [13, 16, 26, 28]).

As the recorded tracks may contain more than just one trip, the tracks are segmented to single trips in a second step. Stays or long stops at activity locations (e.g. office, shopping location etc.) are mostly identified throughout low speeds, like in Axhausen and Schüssler [1] or Menghini et al. [23]. There is also a number of studies which use data preprocessing in a very sparse way – or at least they do not describe it sufficiently (see for example the studies of [18, 27, 34]).

In most studies, there is a third step, which treats the recognition of the transportation mode. However, transport mode recognition methods are mostly not used focusing on cycling as a mode of transport [4, 14, 24, 29]. Furthermore, mode recognition is achieved using different methods (e.g. simple filtering, heuristic or machine learning methods) and speed as the main input parameter.

Apart from mode recognition and trip segmentation, every type of data filtering or data treatment observed in previous studies that uses hard coded thresholds is likely to eliminate many data points or whole tracks, which should not be excluded from the research data set. A very basic speed threshold, for example, fails to detect weather the reason for stopping is a traffic light or an activity, which reduces the cyclists speed. If data is excluded from further data treatment this way, it is lost for further steps like map matching or mode recognition. This can cause a worse overall result in data treatment.

The accuracy of mode recognition raised over the last years. Chung and Shalaby achieved an accuracy of 75% in 2005 realizing mode recognition after the map-matching process and using a threshold for cycling speed. The database was comparatively low, as they used only four bike rides of the same person [9]. In the same year, Stopher et al. [33] used a similar approach reaching an accuracy of 72% of correctly identified bike rides. They therefore combined GIS data (bus stops) with simple speed thresholds. However, the size of the dataset remains unclear [33]. Bohte and Maat [4] also used speed thresholds and applied a decision tree reaching 72% accuracy. In 2010 Reddy et al. [25] first used artificial intelligence (AI) to identify bike trips. They reached an accuracy of 88% of correctly identified trips [25], whereas Gong et al. [14] and Zhang et al. [42] and others did not include cycling as mode in their approach [14, 30, 42]. In contrast, Stenneth et al. [31] focused on the identification of bicycle trips using GIS and GPS data. They applied machine learning algorithms and reached a rate of 89% of correct identified trips. Following the approach of using machine learning (ML), the identification rate ranged from 82% to 100% using sensor data fusion of GPS, acceleration, magnetometer and further sensors [6, 10, 40, 41, 43]. Machine Learning approaches have in common to abstain from using GIS information; because of the problematic data treatment, furthermore they are highly accurate but hard to explain in their classification approaches. Zhang et al. [42] used a two stage approach, which firstly identified active modes of transport using heuristics (decision tree) followed by a ML approach (support vector machine) classifying the other modes. They scored for 95% accuracy but used 19 bike trips, only [42].

For the identification of bicycle trips, the utilisation of heuristics instead of machine learning methods seems to be sufficient and comparatively easy and with less data requirements than ML approaches. It has to be mentioned that the comparability of the referred work is relatively low, as authors of other studies do not state how validation was done. Furthermore, the size of the dataset is even not reported, in some cases. Another problem appears in the way trip segmentation and mode recognition were treated as separate steps of work: a high percentage of correct mode recognition (which is reported) in combination with a lower percentage of trip segmentation is leading to significantly lower values of correct trips [24].

The sections above illustrate that there are different methods that have been applied to process GPS data.

However, there are only few studies that combine and adjust the different steps of GPS data processing specifically to bicycle traffic. Menghini et al. [23] illustrates the existing research gap perfectly as they refer to the "the most probable mode bicycle" using bicycle unspecific pre-processing from Schüssler and Axhausen ([23], p.5). Another example for the need of a bicycle specific pre-processing is the work of Ton et al. [35]. They developed a route choice model and refer to van de Coevering et al. [37] for detailed description of data pre-processing. Gaining insights into the work of van de Coevering et al. [37] reveals that the pre-processing consists of anonymization (cutting of distances track start and end), excluding short routes (< 500 m) and a static trip segmentation (when cyclists stay more than 180 s. within a radius of 300 m). A mode detection has not been implemented or rather described.

It can be summarized, that there are neither well working heuristics nor ML approaches for the pre-processing of bicycle data that comply with central requirements, which are: (a) reducing the used data because a high amount of (different) data can hardly be handled. Furthermore, (b) the applied methods need to be easy to use for practitioners and (c) methods should reveal a high accuracy in terms of bicycle trip recognition. All researched models have in common that validation remains somehow blurred or the number of used bicycle trips for model development is comparatively low. Therefore, the presented contribution reveals a bicycle specific approach, which is based on a large dataset and clear validation criteria.

3 Methodology

To overcome the major shortcomings described in the previous section, we present an approach for bicycle-specific data processing of smartphone-based and crowd-sourced GPS track data. The multi-level approach represents a comprehensive method of mobility data processing focussing on bicycle transport planning.

The data processing method has been developed in two steps. At first, we designed a prototype based on a small sample of labelled track data, which includes all modes of transport and some activities ($n = 49$). We secondly applied the method to a data set containing more than 8900 GPS tracks that have been recorded within the scope of a large bicycle research project. Figure 1 gives an overview over the different levels of the data processing approach, which will be described in the following sections in more detail.

We evaluated the results after applying the data processing approach, varied the model parameters and processed the data again in a last step to get the final results regarding model accuracy.

3.1 Data collection

The data used in this approach has been recorded in the city of Dresden (Germany). The city is located in eastern Germany and is the capital of the federal state of Saxony. About 560,000 inhabitants are living within the cities borders (328.8 km^2) and the mode share of cyclists is about 16% [12].

The used data has been recorded between March and June 2018. In a first step, a small data sample containing 49 tracks of different modes of transport have been recorded and manually labelled by research group members. They contained tracks from car trips (4) as well as tram (5), walking (17), train (1), cycling (15) and sports cycling (4) trips. Trip length ranged between ten minutes (walking) and 3 hours (sports cycling). The data was analyzed and used to derive a first data processing approach. The labelled test data were analysed using a spreadsheet (MS Excel).

To evaluate and improve the developed approach another data set was used. The second data set was recorded by 187 volunteer cyclists, which participated in a bicycle research project covering the city area of Dresden (Germany). The participants were selected out of 10,000 people taking part in an online survey to determine types of cyclists [11]. The data set has also been collected in 2018 (June/July). The sample consists of 80 female and 100 male riders aging from 16 to 88 years. We used the Cyface smartphone application (for iOS and Android) for data collection, which provided the possibility to import the data manually and automatically. Data collection was performed with a frequency of 1 Hz, which was the case for all values 99.7%. The transferred data contained information regarding latitude (lat), longitude (lon), speed and accuracy. The cyclists recorded 8909 measurements resulting in 5300

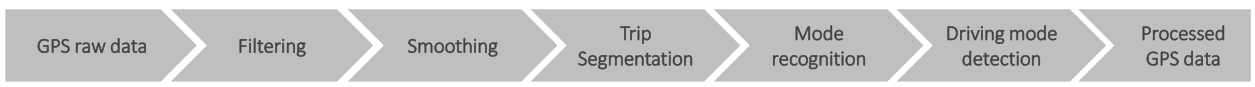

Fig. 1 Steps of data processing

valid bicycle trips after data preprocessing. Mean speed over all recorded measurements in the unprocessed dataset was 3.1 m/s. Mean speed over all processed bicycle trips was around 3.96 m/s [21].

3.2 Data import

The data recorded by cyclists using the Cyface smartphone application during the research project was automatically transferred from the study participants' smartphone to a PostgreSQL database on a SSH secured virtual machine in the universities net via WLAN connection. The data contained latitude, longitude, signal accuracy and speed values. Due to technical issues, not all GPS points were in correct chronologic order regarding so that the indexing, which was initially done be the app, was rearranged to gain the proper time series of GPS point within a track.

3.3 Filtering

In a first and very basic step of data processing, all tracks with a timespan less than 30 s are eliminated because these tracks are likely to not contain any reasonable information. We experienced that such data occur because people often tested the record function of the application, initially. A further filter treats GPS leaps and excludes GPS points in the track data whose speed is higher than 25 m/s (90 km/h) – a speed, which is not very likely to be reached, even not by motorized vehicles on inner city highways. This threshold can be varied for different data processing tasks as it is not hard coded. The goal is only to eliminate data points, which are very unlikely to occur because of cycling behaviour or to be recorded from cyclists. Speed is calculated in a further step because calculation methods for GPS speeds differ between different smartphone (or rather software) and sensor manufacturers, which could lead to inequalities in data. The calculation follows

$$v = \left(\frac{S_{i-i+1}}{t_{i+1} - t_i}\right) \text{ and } s_{i-i+1} = \sum_i^n Hav_{i-i+1} \tag{1}$$

Whereas v is the speed calculated with distance s between timestamp t_i and t_{i+1} of the GPS point i and i + 1.The distance s is determined calculating the haversine distance following eq. (2).

With *Hav* as the haversine distance between GPS points P_i and P_{i+1} and Lat/Lon as the coordinates in terms of latitude/longitude of the referring points.

Another filter treats GPS accuracy. If accuracy is below 50 m, which means that there are GPS points with a potential error bigger than a diameter of 50 m around the original position, GPS points are excluded. This is an effective way to react on GPS errors such as reflection or signal refraction like the multipath problem.

3.4 Smoothing

Calculated raw speeds values show highly erratic gradients so that data smoothing is essential for the next steps of data processing. We use an already established method called *Gaussian Smoothing* for that [26]. In contrast to a sliding average, the degree of smoothing is weighted by the distance between the processed point and all other points within a 15-s time window using Gaussian distribution. The calculation follows

$$\tilde{c}(t) = \frac{\sum_j (w(t_j) * c(t_j))}{\sum_j w(t_j)} \tag{3}$$

with

$$w(t_j) = \exp - \frac{(t - t_j)2}{2\sigma 2} \tag{4}$$

$\sigma = 10$Parameter σ represents the kernel bandwidth, which is set at 10 s ($\sigma = 10$) similar to previous research [1]. $\tilde{C}$ is the resulting smoothed speed value of the GPS point at i at time t. C is the raw speed value at the GPS points at j at time t_j. A further dynamic filter is implemented, which reduces the span of smoothing when the smoothing window is running towards a stop (for example at a traffic light) to preserve the sharpness of the original data instead of overwriting it.

3.5 Trip segmentation

The trip segmentation identifies the actual trips within an uploaded GPS track. This step is necessary because one uploaded GPS track may contain several trips (e.g. when a user keeps on tracking while doing paper work in the office). Thus, the tracks have to be evaluated and, if necessary, segmented into several parts (e.g. bike – office – bike). Different variables, such as speed, track point density or directional change can be used for the evaluation. Gong et al. [14] give an overview over useful variables that can be considered for the segmentation. Further variables and methods can be found in Kohla

$$\text{Hav}_{i-i+1} = 2 * R_{\text{Earth}} * \sin^{-1} \sqrt{\sin\left(\frac{\text{Lat}_{i+1} - \text{Lat}_i}{2}\right)^2 + \text{cosLat}_i * \text{cosLat}_{i+1} * \sin\left(\frac{\text{Lon}_{i+1} - \text{Lon}_i}{2}\right)^2} \tag{2}$$

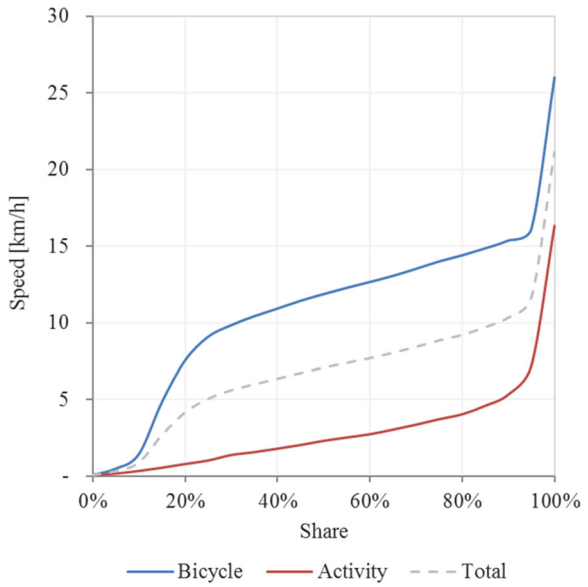

Fig. 2 Example of differing speed distribution

[19, 20], Schüssler & Axhausen [26], Zong [43] Zhang et al. [42] Biljecki [2, 3], and Shen & Stopher [28].

The developed trip segmentation algorithm determines for each GPS point whether it belongs to a trip (e.g. cycling) or to an activity at a location (e.g. office). The developed algorithm therefore comprises speed, travelled distance and changes in the directional vectors for each individual point of a track. These are proper indicators to detect activities due to several reasons:

1. Speed value distribution varies significantly between a stay at a location and a ride. Figure 2 exemplifies the speed distribution of a typical bicycle trip and an uploaded track of an activity from the labelled test data set.
2. Staying at a location often causes signal loss or interferences through shield and reflection effects caused by walls etc.. This leads to hopping of GPS points and, thus, disproportionate change of directional vectors. Figure 3 exemplifies the change

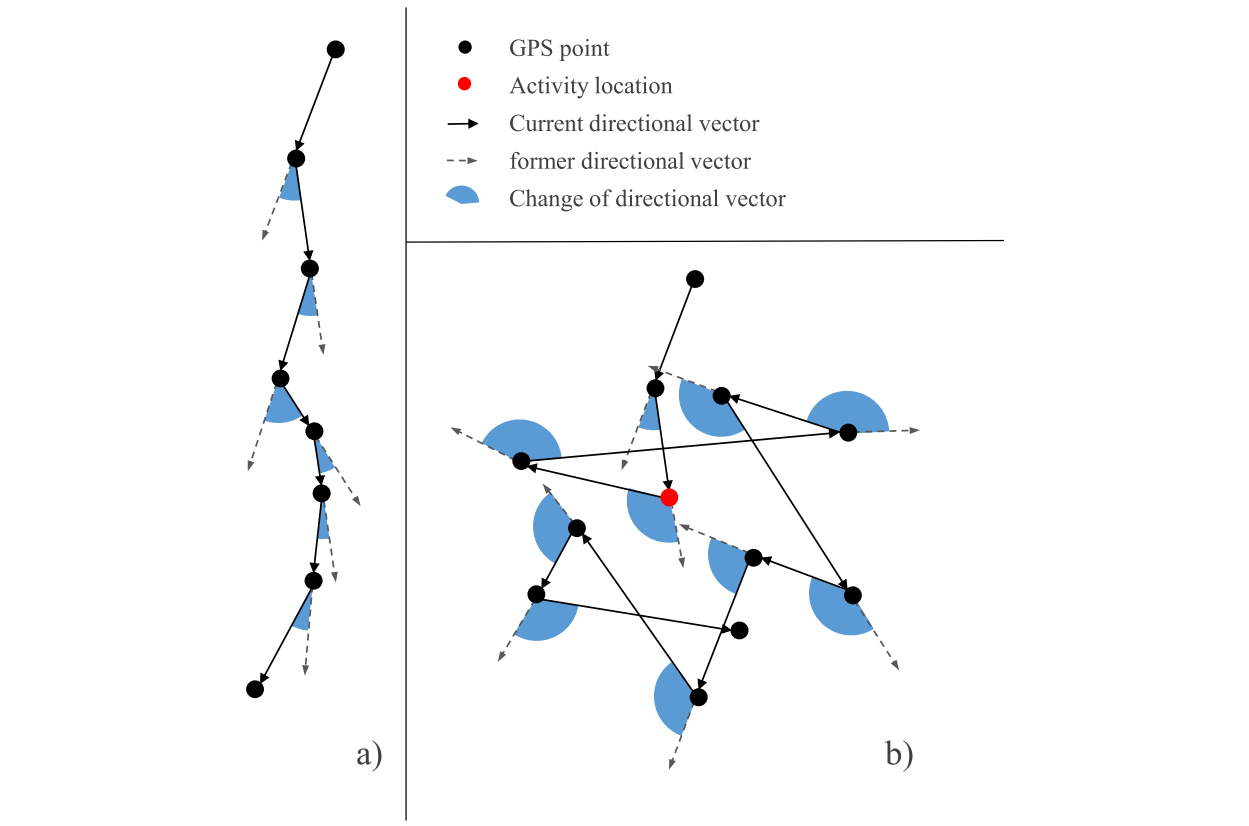

Fig. 3 Change of directional vectors

of the directional vectors while cycling (3a) and for activities at a location (3b)
3. In contrast to a trip there is only little gain in distance while staying at a location (distance gain caused by GPS point hopping is significantly lower than distance covered while cycling). This, of course, comes along with lower speeds (see 1.).

As the input variable may vary significantly and the variation may occur during activities as well as during trips, a new variable τ is computed to improve the identification of activity points. The variable τ combines the mentioned input variables and is calculated as followed:

$$\tau_i = v_i * d_i * \frac{1}{\Delta head_i} \tag{5}$$

with v_i as smoothed speed at point i, d_i as smoothed distance at point i and $\Delta head_i$ as change of the directional vector at point i. Thus, the variable τ_i (tau) improves the determination through exaggerating.

The smoothed speed s_i as well as the smoothed distance d_i at point i are results of former calculations. The change of the directional vector at point i is calculated following the equations

$$\Delta head_i = \frac{1}{n} \sum_{k=-7}^{n} \Delta head_i \tag{6}$$

and

$$\Delta head_i = \begin{cases} |\dot{\alpha}_l - \dot{\alpha}_j| & as \Delta\dot{\alpha}_{ij} \leq 180 \\ ||\dot{\alpha}_l - \dot{\alpha}_j|, -, 360| & as \Delta\dot{\alpha}_{lj} > 180 \end{cases} \tag{7}$$

and

$$\dot{a}_{lj} \begin{cases} 90 - \alpha & , as \alpha > 0 \cap \alpha \leq 90 \\ 360 - (\alpha - 90) & , as \alpha > 90 \cap \alpha \leq 180 \\ \sqrt[2]{(a-90)^2} & , as \dot{\alpha} \leq 0 \cap \alpha \geq -90 \cup \alpha < -90 \cap \alpha \geq -180 \end{cases} \tag{8}$$

and

$$\alpha = ARCTAN2\,(x, y) * \left(\frac{180}{\pi}\right) \tag{9}$$

and

$$x = \cos\left(lat_j * \frac{\pi}{180}\right) * \sin\left(lon_j * \frac{\pi}{180}\right) - lon_i * \frac{\pi}{180} \tag{10}$$

and

$$y = \cos\left(lat_i * \frac{\pi}{180}\right) * \sin\left(lat_j * \frac{\pi}{180}\right) - \sin\left(lat_i * \frac{\pi}{180}\right) * \cos\left(lat_j * \frac{\pi}{180}\right) * \cos\left(\left(lon_j * \frac{\pi}{180}\right) - \left(lon_i * \frac{\pi}{180}\right)\right) \tag{11}$$

Following the calculation, there finally is a τ for each GPS point of an uploaded track. To determine if a GPS point is part of a trip or part of an activity at a location the algorithm checks the τ value for each point within a gliding time window of 180 s. This value is then checked against a modifiable threshold value. If τ of a point within the time window is < 1.5 the point is identified as part of a trip (otherwise as part of an activity). The τ value has been derived from data analysis using the test data set (see section 3). The algorithm developed here uses a time window of 180 s. Other approaches also used a gliding time window for determination but the length of windows varies over the different studies between 120 s (see for example [22, 33, 36, 38]) and up to 300 s (see for instance Wolf et al. 2004 [4, 14, 39];). For the event of a lost GPS signal, we decided to keep the time window and split trips after 180 s without data input. If the signal loss is shorter, we keep the methodology of averaging the existing τ values, following the thesis, that in doubt a standstill is less likely than a continuing ride.

Figure 4 illustrates how speeds, changes of the heading of the directional vector as well as the covered distance develop over time. It reveals that using the new parameter τ is by far more adequate to identify if a part of the GPS tracks belongs to a trip or an activity. Trip and activity sections are displayed in red (1 = trip, 0 = activity). The dashed line represents the transition from trip to activity in this example. Functionality and validity is presented in section 4.

3.6 Mode recognition

In order to determine the used mode of transport, a rule-based heuristic mode recognition model has been developed based on in-depth GPD data analysis and existing approaches. We implemented the developed heuristic as a decision tree with different decision levels. Passing the model, each trip is assigned to one of four possible traffic modes (walk, bicycle, leisure bicycle, other). In the following sections, we describe the input data, the processing as well as the output of the model.

To determine the transport mode, the model needs mode-specific values that represent the characteristics of the used mode. These values have been identified by analysing labelled tracks (known mode of transport) from the test data set (see section 3). The analysis revealed that there are different values, which may be properly used for mode recognition, such as different

Fig. 4 Illustration of speed, heading, distance and TAU values over a trip and activity section

acceleration and speed percentiles, distances or detour factors. Figure 5 shows examples of speed distributions (a) and distribution of acceleration (b) of different modes of transport.

The figure reveals that speed distribution (Fig. 5a) significantly differs between different modes and, thus, is useful for mode recognition. The mode 'walk', for example, shows very low speeds (80% percentile < 10 km/h) whereas speed of 'bicycle'is significantly higher in the first percentiles (e.g. 30% percentile > 10 km/h). On the other hand, the distribution of acceleration does not show distinct characteristics. They, therefore, can hardly

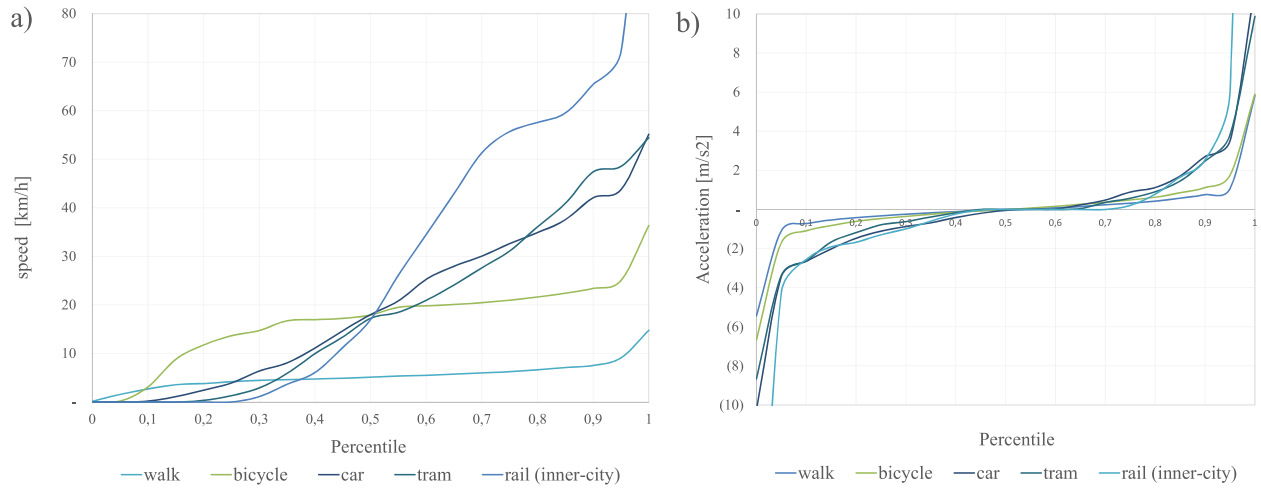

Fig. 5 Distribution of **a**) speed and **b**) acceleration of different modes of transport

be used for mode recognition – especially because minimum and maximum valued depend on GPS signal quality. Errors occur through reflection or signal loss and, thus, influence acceleration and maximum speeds. Hence, high (e.g. 90%) or low (e.g. 10%) percentile should be used for mode recognition, only.

A further and very important variable is the distance and the detour factor. Distance and detour can significantly improve mode recognition because the distances covered by different modes and the corresponding detour differ considerably. Table 1 shows some examples of detour factors for different modes.

The detour factor is especially important to detect e.g. leisure and sport trips of cyclist but it can also be useful to distinguish e.g. between rail and other modes, as its detour factor is normally very low. As we are not considering road networks or other GIS data, the detour factor is calculated as the beeline. This is mandatory, because using shortest paths instead would mean to make a pre assumption for a specific mode for each trip because searching the shortest route is restricted by using mode-specific (allowed) infrastructure.

A decision tree is used to test and implement the mode recognition model. The rules implemented at each node of the tree are based on the results of data analysis. The implemented multi-level decision tree contains three decision levels. The main input is the calculated values for the following variables:

- 20% percentile of speed of a trip
- 80% percentile of speed of a trip
- 90% percentile of speed of a trip
- Trip distance
- Detour factors of a trip

The mentioned values of the variable are calculated for each trip. The computation of the n-% percentile executed by outputting the value at point n-% of the ordered set of values. Trip distance is computed cumulating the distances between the GPS points as

$$d_{trip} = \sum_{x=1}^{n} s_{i-i+1} \tag{12}$$

with d_{trip} as trip distance and s_{i-i+1} as distance between two points P_i und P_{i+1} of a trip. The detour factor of a trip is computed using the trip distance and the beeline (distance between the first and the last GPS point). The calculation is carried out as

$$DF_{trip} = \frac{d_{trip}}{dl_{trip}} \tag{13}$$

with DF_{trip} as detour factor of a trip and dl_{trip} as beeline between first and last point of a trip. The computation of dl_{trip} is carried out following the calculation of *Hav* (see section 3). If original coordinates of the points are used, they have to be converted from degree to radian measure following

$$Lat_{RAD} = Lat_{DEG} * \frac{\pi}{180} \tag{14}$$

As all values for the presented variable are calculated, they can be used for the mode recognition decision tree model. Starting with the root node, the calculates attributes/values of each trip are checked at every node of the decision tree. Figure 6 illustrates the used decision tree for mode recognition.

The modes are defined as follows:

- Mode 1: walk
- Mode 2: bicycle (leisure)
- Mode 3: bicycle
- Mode 4: other

The decision rules at each node consider speed values, distances and detour factors of each trip. The rules or requests at each node are defined as:

1. *if A = true, then M* = 1, *else* 2.
2. *if* $B \cap (C \cup D) \cap F = true$, *then M* = 2, *else* 3.
3. *if E = true, then M* = 3, *else M* = 4

with:

$A = v80_{trip} \leq \alpha_{M1}$
$B = v20_{trip} \geq \alpha_{M2.1}$
$C = D_{trip} > \delta_{M2}$
$D = DF_{trip} > \gamma_{M2}$
$E = v90_{trip} \leq \alpha_{M3}$
$F = v80_{trip} < \alpha_{M2.2}$

and:
$v80_{trip}$ – 80%-percentile of trip speed.
α_{M1} – Threshold for the 80%-percentile for mode 1.
$v20_{trip}$ – 20% percentile of trip speed.
$\alpha_{M2.1}$ – Treshold for the 20%-percentile for mode 2.
$\alpha_{M2.2}$ – Treshold for the 80%-percentile for mode 2.
D_{trip} – Trip distance.
δ_{M2} – Treshold for trip distance of mode 2.
DF_{trip} – Detour factor of the trip.
γ_{M2} – Treshold for the detour facor for mode 2.
$v90_{trip}$ – 90%-percentile of trip speed.

Table 1 Examples of typical detour factors from the data set

Mode	Detour factor
Bicycle	1,4
Bicycle (leisure)	24,3
Walk	1,4
Train	1,1

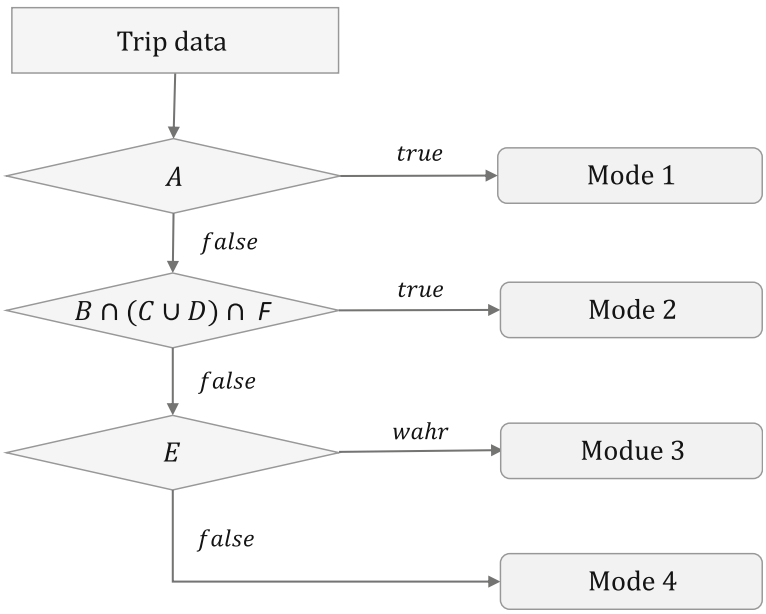

Fig. 6 Decision tree for mode recognition

α_{M3} – Treshold for the 90%-percentile for mode 3.

The following values (see Table 2), which have been identified in the analysis of the trips data, are input parameters for the decision tree.

The recognition of further modes is, of course, possible and feasible. However, the focus of the investigation is not the perfect mode recognition of all existing modes. Therefore, an aggregation of modes like car or bus to mode 4 (other) seems to be viable. The result of the decision tree is an assigned mode for each trip passing it. Accuracy of the method is described in chapter 4.

3.7 Driving mode detection

In order to provide information about the cycling behaviour, this step determines the driving mode of cyclists during the trip. The model distinguishes between four different driving modes (stop, acceleration, constant movement, deceleration). The driving mode is mainly defined by speed (constant movement vs. no movement) and acceleration (acceleration vs. deceleration), which is calculated for each GPS point of a trip. Mode detection is essential to detect real acceleration and deceleration processes. The following equations were used to define the four driving modes:

$$if \;\; MAX\left(v_i \ldots v_j\right) < 0,2\frac{m}{s} \; then'' \text{Stop}'' \tag{15}$$

with

$$i = 1 \; and \; j = 180 \tag{16}$$

and

$$if a_i < \; -0,2\frac{m}{s^2} then \tag{16}$$

and

$$if \; a_i > 0,2\frac{m}{s^2} \; then'' \text{acceleration}'' \tag{17}$$

and

$$else \;\; \text{“constant movement”} \tag{18}$$

Driving mode detection can be seen as part of the data pre-processing for further research items like specifying the behaviour of different types of cyclists.

Table 2 parameter threshold values used in pre processing

Threshold	Value	Unit
α_{M1}	8.0	km/h
$\alpha_{M2.1}$	10.0	km/h
$\alpha_{M2.2}$	50.0	km/h
δ_{M2}	30.0	Km
γ_{M2}	3.0	–
α_{M3}	40.0	km/h

3.8 Validation

To validate the data processing, systematic and random errors need to be assessed. Systematic errors are part of the data collection and processing, whereas random errors occur more often on the participants side of a study. As this paper is about the data processing, we mainly focus on systematic errors.

We used initial model parameters derived from data analysis to test and validate the model for the GPS cycling data set (see sections above). The validation of the model results was then carried out in two steps. We firstly examined about 150 tracks that

Table 3 Altering of parameter combinations for different steps of data pre processing

Version	Speed [m/s] Upper Boundary	Speed [m/s] Lower Boundary	T (tau)	Speed "Stop" Threshold [m/s]	Time Gap [s]	Acc. Threshold [m/s^2]	v20-percentile ($\alpha_{M2.1}$) [km/h]	v80-percentile ($\alpha_{M2.2}$) [km/h]	v80-percentile (α_{M1}) [km/h]	v90-percentile (α_{M3}) [km/h]	Distance (δ_{M2}) [km]	Detour Factor (γ_{M2})
1	25	0,2	3	0,2	60	0,1	10	–	8	30	30	3
2	25	0,2	3	0,2	60	0,1	10	–	8	35	30	3
3	25	0,2	3	0,2	60	0,1	10	–	8	40	30	3
4	25	0,2	1	0,2	180	0,2	10	–	8	40	30	3
5	25	0,2	3,5	0,2	180	0,2	10	–	8	40	30	3
6	25	0,2	3	0,2	180	0,2	10	–	8	40	30	3
7	25	0,2	3	0,2	180	0,2	10	50	8	40	30	3

showed striking characteristics (e.g. very high speeds, very long or short tracks or tracks with few points). Secondly, we selected 150 tracks randomly for validation. Additionally, we examined the track samples visually to evaluate the performance of trip segmentation and mode recognition. We used QGIS and a Python plug-in for the last step.

After the evaluation of one combination of parameters, we altered the parameters with impact on segmentation or mode recognition. After verifying the combination for a couple of tracks using the Python plug-in, we ran the data pre-processing on the whole data set again and restarted the evaluation for another 150 randomly selected tracks. Table 3 shows the different parameter combinations of our iterative pre-processing attempts. The percentage of "valid trips" shows the value of trajectories, which are not marked as activities and are not discarded due to little amounts of time or GPS-points.

In addition to potential errors in the data processing, the dataset had to be checked for systematic errors in data collection. A number of errors occurred in the data regarding the different smartphone devices. As they could result from hardware specifications as well as software issues, we checked for high amounts of segmented trips after data pre-processing and additionally for a low number of GPS points compared with the trip length. A big number of trips is an indicator for a) a very ambitious cyclist or b) a smartphone, whose recorded tracks caused problems in the trip segmentation step due to non-adjacent GPS-trajectories. However, a very low number of GPS points in combination with great trip length is an indicator for power saving modes. This shut down in data collection occurs when the operating system software of the smartphone moves the activity of the application to the background, which disables the app to collect data. Both kind of errors may not precisely be linked with a special smartphone type or model. There was a slight tendency for Chinese fabricates (e.g. Huawei and Xiaomi), which was not significant so far. Corrupt data sets were deleted after the validation process.

4 Results

The result of the data processing model is supplementary information for each GPS point within a track: whether it belongs to an activity or a trip, the corresponding mode of transport and driving mode). Figure 7 illustrates how real behaviour and the referring trip and mode information within the processed data look like.

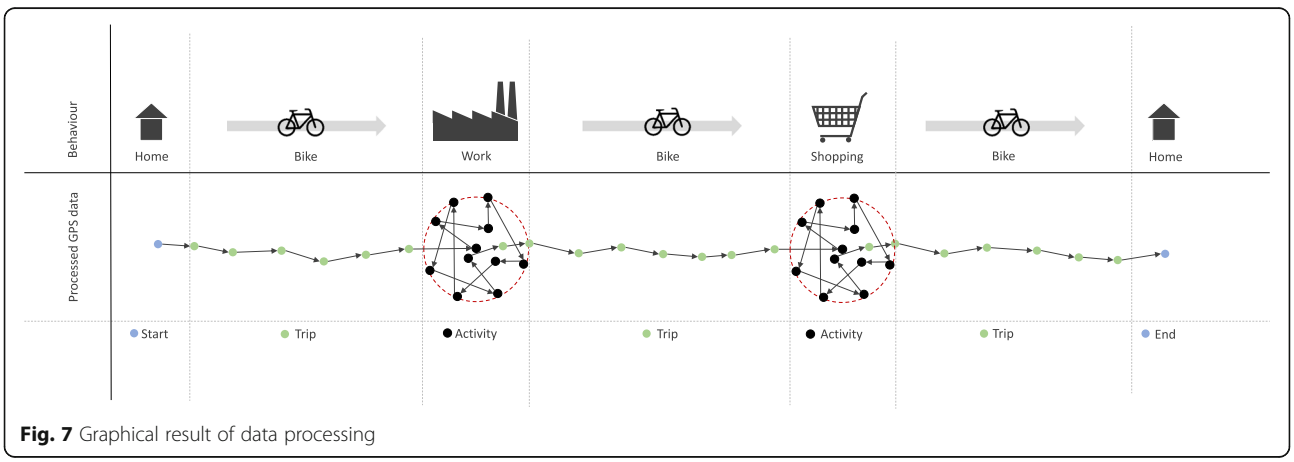

Fig. 7 Graphical result of data processing

The validation of the trip segmentation model shows high accuracy, over all. Using adapted parameters derived from the first iteration, about 82% of the tracks of the GPS bicycle data sample were segmented 100% correctly in the second iteration (see Fig. 8). Only about 18% of the tracks have not been segmented correctly (15% of them segmented once and 3% segmented several times). However, there is huge potential to diminish the remaining errors through further model adaptations. Doing further iterations and adapting the gliding time window (especially widening it) or adapting the τ value may help to increase the accuracy of the trip segmentation up to around 95% (see Fig. 8).

The validation of the transport mode recognition revealed that the initial calculation already produced good results. About 85.4% of the transport modes were correctly classified. Around 7.8% of the trips were incorrectly classified due to errors in the trip segmentation (partially correct) and only 6.8% were classified incorrectly because of inaccuracy of the mode recognition model (see Fig. 9). After altering model parameters nearly 87.8% of all cycling trips in the sample were detected correctly. Only about 12% of the trips have not been assigned correctly. The incorrect classification are partially (about 4.9%) traced back to errors within the trip segmentation because trip length, for instance, indirectly influences transport mode recognition.

However, the overall results can be assessed as good. There, furthermore, is huge potential to improve the mode recognition. An improvement of trip segmentation, for instance, directly affects the accuracy of the mode recognition – this close connection has already been identified between the different iterations (see Fig. 9). The potential through trip segmentation improvements is around 3.9%. Thus, a total accuracy of 91.7% can be achieved (potential 1). Further improvement can also be achieved adapting the parameters of the transport mode recognition model itself. A revision of the model itself or an adaptation of model parameters could lead to an improvement of 8.3%, which finally results in an accuracy of 96.1% (see Fig. 9, potential 2).

The model accuracy and the different potentials were cross validated using another sub sample of the dataset after the parameter fitting to avoid overfitting.

5 Discussion

The presented data processing model represents an adequate approach to overcome the still existing weakness in GPS-based cycling data (see section 0 and 13). It therefore represents an important contribution in the field of bicycle-specific data processing.

Since trip segmentation and transport mode recognition show high accuracies (81% and 88%), the overall results of the data processing model can be assessed as pretty good – especially because the presented model is based on GPS data, only. Additionally it has to be mentioned, that we trained our model with complex intermodal mobility chains, where walking trips (mostly under 180 s.) at the begin or end of bicycle trips where the main cause for incorrect segmentation or mode detection.

Although there are still incorrectly segmented tracks and wrongly recognized transport modes, there is a high potential to eliminate the remaining errors by further model adjustments. Implementing appropriate measures, we found that the trip segmentation accuracy could be further increased up to approximately 95% by adapting model parameters. Comparing the results from the initial calculation with the results from the second iteration also revealed that an improved trip segmentation directly effects the transport mode recognition. We found that improving the trip segmentation could increase the

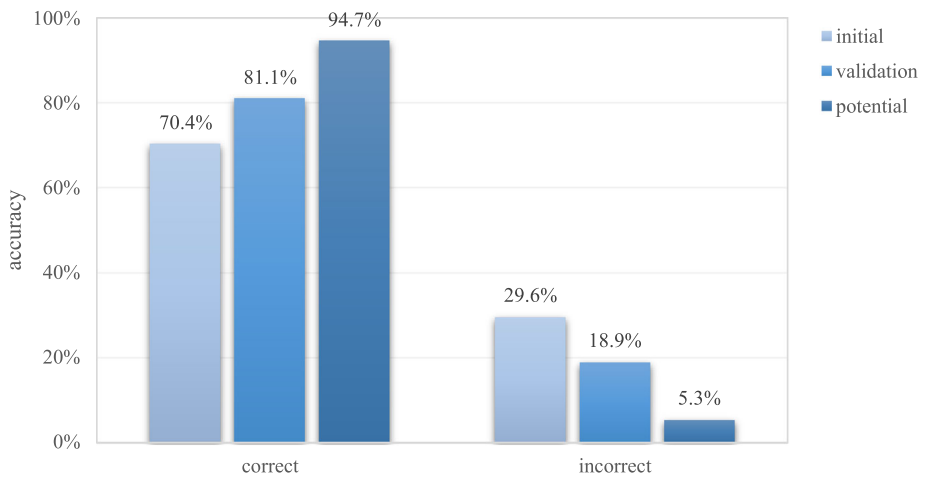

Fig. 8 Accuracy and potential accuracy of trip segmentation

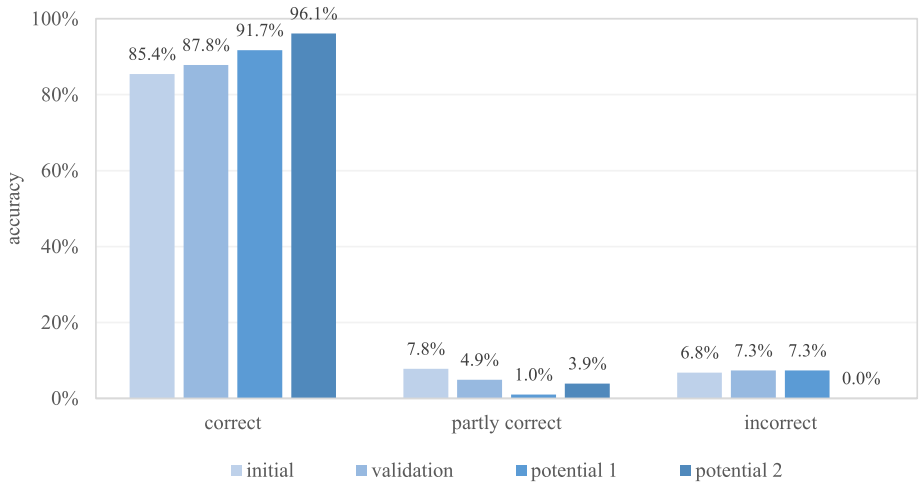

Fig. 9 Accuracy and potential accuracy of mode recognition

accuracy of transport mode recognition up to 91.7%. Adjustments of model parameters of the transport mode recognition can further lead to improvement. We found that adapting the parameters could lead to an overall accuracy of 96.1%.

Comparing the potential loss of data (3.9%) with the loss when simple threshold-based filtering models are applied, illustrates the potential of the presented approach. From 11.397 million GPS points in the raw data set, the initial data processing led to around 11 million remaining points (3% loss in data). In comparison to the used approach, a threshold filtering with a lower boundary of 0.1 m/s and an upper boundary of 10 m/s resulted in 7.7 mil data points and a loss of 32% of information without knowing the mode and trip length of the processed data. Introducing a filter of 1.4 m/s (5 km/h; walking speed) without an upper boundary results in 6.945 mil data points and therefor a loss of 39% of the initial data while excluding walking (almost completely), activities and stops of cyclists. On the other hand, our preprocessed data contained only 7.35 mil. Data points which were valid and had the mode "bike" or "sports bike". It can be considered, that there is an amount of about 30% noisy data, which has to be filtered. However, our processed bike trips contained more than 892,000 data points with a speed lower than 0.1 m/s, which is about 11% of the whole bicycle dataset that would be lost using a 0.1 m/s threshold.

According to model accuracy can be stated that in comparison to the reviewed studies (see section 2), the presented model shows a high accuracy while using few data or rather no further data than GPS. At the same time, the used heuristic represents a transparent and comprehensible, which is easy to implement. Other studies using heuristics, such as the ones of Chung & Shalaby [9], Bohte & Maas [4] or Stopher et al. [32] for instance, show least accuracy values of all studies (between 72% and 75%) in identifying bicycle trips, although they used further information (GIS data). In contrast, the heuristic developed by Zhang et al. [42] reveals a high accuracy (95%) using GPS data only. They identify cycling trips by taking into account values of speed, acceleration and heading. However, Zhang et al. [42] consider 19 bike trips, only, which is very little. Furthermore, the study aimed on identifying car trips using machine-learning methods and the bicycle mode detection was achieved in a upstream step. There is no further information regarding model accuracy for the other studies using heuristics (e.g. [14] and [30]).

Other studies use different methods (machine learning) and different data. The studies, which use machine-learning methods generally achieve higher accuracy values (between 82% and 93%) than models that use heuristics (e.g. [6, 7, 10, 31, 41]). However, the mentioned studies use further data such as GIS data or further data from smartphone sensors (e.g. accelerometer, gyroscope or magnetometer). This increases data processing complexity and, therefore, hampers model implementation. Studies using ML, which do not use further data for bicycle trip identification show accuracy values between 88% (e.g. [25, 40]) and 100% [43]. However, the underlying machine learning models are very complex and can hardly be reproduced (see for instance [43]), which hampers model implementation and utilization, especially for practitioners. Additionally small homogeneous datasets can cause model overfitting.

6 Conclusion

Aim of the study was to develop a bicycle specific data processing approach, which is capable to process big GPS data sets and easy to use and to implement for practitioners. The goal was to create a method which is

highly transparent, flexible and interpretable (no black box). Furthermore a high accuracy was an essential requirement. The study results can be compared with other studies that focussed on using and identifying bicycle trips (see section 2). The most important criteria for comparison are a) the used data, b) the developed model and c) the accuracy of bicycle trip identification. Comparing these parameters is important towards implementing a manageable model (criterion b), which is able to identify bicycle trips with little or no further input data than GPS tracks (criterion a). It is further supposed to show a high accuracy (criterion c) to assure a minimum loss of bicycle trips for further data analysis.

We summarize that the developed data processing model generally represents an adequate approach to overcome the gaps in bicycle data processing – especially because it represents a simple but robust approach that is easy to implement and has low data requirements. In contrast to other approaches the method can be considered as novel at the data filtering stage because data loss can be reduced effectively. Furthermore the model is very flexible because key values can be adopted to different context. The focus on bicycle transport and the bicycle specific thresholds contributes to research in this field. However, there is still potential for improvement achieved by both, smaller model adaptations and applying other classification methods.

Acknowledgements
Not applicable.

Authors' contributions
Sven Lißner and Stefan Huber both contributed likewise to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript. The authors read and approved the final manuscript.

Funding
The research was funded by the initiative mFund through the federal ministry of transportation (BMVI).
Open Access Funding by the publication Fund of the TU Dresden. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials
Raw data is not publically available due to privacy issues.

Competing interests
There are no financial and non-financial competing interests.

Received: 15 April 2020 Accepted: 4 December 2020
Published online: 13 January 2021

References

1. Axhausen, K. W., & Schüssler, N. (2008). *Identifying trips and activities and their characteristics from GPS raw data without further information Conf Pap*.
2. Biljecki, F. (2010). Automatic segmentation and classification of movement trajectories for transportation modes. *Master Thesis, Delft University of Technology*. https://doi.org/10.4233/uuid:654587d2-6e93-4619-ab9a-29d95f843f35.
3. Biljecki, F., Ledoux, H., & van Oosterom, P. (2013). Transportation mode-based segmentation and classification of movement trajectories. *International Journal of Geographical Information Science*, *27*(2), 385–407 https://doi.org/10.1080/13658816.2012.692791.
4. Bohte, W., & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, *17*(3), 285–297 https://doi.org/10.1016/j.trc.2008.11.004.
5. Broach, J., Dill, J., & Gliebe, J. (2012). Where do cyclists ride? A route choice model developed with revealed preference GPS data. *Transportation Research Part A: Policy and Practice*, *46*(10), 1730–1740 https://doi.org/10.1016/j.tra.2012.07.005.
6. Byon, Y. J., & Liang, S. (2014). Real-time transportation mode detection using smartphones and artificial neural networks: Performance comparisons between smartphones and conventional global positioning system sensors. *Journal of Intelligent Tranportation Systems: Technology, Planning, and Operations*, *18*(3), 264–272.
7. Carpineti, C., Lomonaco, V., Bedogni, L., Di Felice, M., & Bononi, L. (2018). *Custom dual transportation mode detection by smartphone devices exploiting sensor diversity*, (pp. 1–15).
8. Charlton, B., Sall, E., Schwartz, M., & Hood, J. (2011). *Bicycle route choice data collection using GPS-enabled smartphones*, *TRB 2011 Annu Meet* (pp. 1–10).
9. Chung, E. H., & Shalaby, A. (2005). A trip reconstruction tool for GPS-based personal travel surveys. *Transportation Planning and Technology*, *28*(5), 381–401 https://doi.org/10.1080/03081060500322599.
10. Dabiri, S., & Heaslip, K. (2018). Inferring transportation modes from GPS trajectories using a convolutional neural network. *Transportation Research Part C: Emerging Technologies*, *86December*, *2017*, 360–371.
11. Francke, A., Anke, J., Lißner, S., Schaefer, L., Becker, T., & Petzoldt, T. (2019). Are you an ambitious cyclist? Results of the cyclist profile questionnaire in Germany. *Traffic Injury Prevention*, *20*(sup3), 10–15.
12. Gerike, R., Hubrich, S., Ließke, F., Wittig, S., & Wittwer, R. (2020). *Sonderauswertung zum Forschungsprojekt 'Mobilität in Städten - SrV 2018'*.
13. Ghanayim, M., & Bekhor, S. (2018). Modelling bicycle route choice using data from a GPS-assisted household survey. *European Journal of Transport and Infrastructure Research*, *18*(2), 158–177.
14. 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*(2), 131–139 https://doi.org/10.1016/j.compenvurbsys.2011.05.003.
15. Harvey, F., & Krizek, K. (2007). Commuter bicyclist behavior and facility disruption. *Transportation Research Board*, *60* https://trid.trb.org/view.aspx?id=811576.
16. Hood, J., Sall, E., & Charlton, B. (2011). A GPS-based bicycle route choice model for San Francisco, California. *Transportation Letters*, *3*(1), 63–75 https://doi.org/10.3328/TL.2011.03.01.63-75. http://www.tandfonline.com/doi/full/10.3328/TL.2011.03.01.63-75.
17. Jestico, B., Nelson, T., & Winters, M. (2016). Mapping ridership using crowdsourced cycling data. *Journal of Transport Geography*, *52*, 90–97 https://doi.org/10.1016/j.jtrangeo.2016.03.006.
18. Jónasson, Á., Eiriksson, H., Eðvarðsson, I., Helgason, K., & Sæmundsson, T. (2013). *Optimizing expenditure on cycling roads using cyclists ' GPS data*, (pp. 1–20) http://trauzti.com/files/urban-routing.pdf.
19. Kohla, B. (2012) *'MODE – Automated Detection of Motorised Transport Modes out of technology-based mobility data' Project report*. https://graz.pure.elsevier.com/de/publications/mode-verfahren-zurautomatisierten-identifikation-motorisierter-v.
20. Kohla, B. (2014). *'Erkennung von Wegetappen und Verkehrsmitteln für Mobilitätserhebungen mit mobilen Erhebungsgeräten', Presenation at the Academia Conference (Hochschultagung), Bad Herrenalb* (vol. 2014).
21. Lißner, S., Huber, S., Lindemann, P., Anke, J., & Francke, A. (2020). GPS-data in bicycle planning: "Which cyclist leaves what kind of traces?" Results of a representative user study in Germany. *Transportation Research Interdisciplinary Perspectives*, *7*, 100192.
22. Menghini, G., Carrasco, N., Schüssler, N., & Axhausen, K. W. (2009). Route choice of cyclists in Zurich: GPS-based discrete choice models. *Arbeitsberichte Verkehrs- und Raumplanung, IVT , ETH Zurich*, *544*(544), 1–25.
23. Menghini, G., Carrasco, N., Schüssler, N., & Axhausen, K. W. (2010). Route choice of cyclists in Zurich. *Transportation Research Part A: Policy and Practice*, *44*(9), 754–765 https://doi.org/10.1016/j.tra.2010.07.008.
24. Prelipcean, A. C., Gidófalvi, G., & Susilo, Y. O. (2017). Transportation mode detection–an in-depth review of applicability and reliability. *Transport Reviews*, *37*(4), 442–464 https://doi.org/10.1080/01441647.2016.1246489.
25. Reddy, S., Shilton, K., Denisov, G., Cenizal, C., Estrin, D., & Srivastava, M. (2010). *Biketastic: Sensing and mapping for better biking*. Proceedings of the

SIGCHI Conference on Human Factors Computer System (CHI '10), 3, 1817–1820. https://doi.org/10.1145/1753326.1753598. http://www.scopus.com/inward/record.url?eid=2-s2.0-77953965556&partnerID=tZOtx3y1%5Cn. http://dl.acm.org/citation.cfm?id=1753326.1753598.

26. Schüssler, N., & Axhausen, K. W. (2008). Identifying trips and activities and their characteristics from GPS raw data without further information. In *ETH Zürich research collection*.
27. Segadilha, A. B. P., & Sanches, S. d. P. (2014). Analysis of bicycle commuter routes using GPSs and GIS. *Procedia - Social and Behavioral Sciences*, *162*, 198–207 https://doi.org/10.1016/j.sbspro.2014.12.200.
28. Shen, L., & Stopher, P. R. (2014). Review of GPS travel survey and GPS data-processing methods. *Transport Reviews*, *34*(3), 316–334 https://doi.org/10.1080/01441647.2014.903530.
29. Shin, D. (2016). Urban sensing by crowdsourcing: Analysing urban trip behaviour in Zurich. *International Journal of Urban and Regional Research*, *40*(5), 1044–1060 https://doi.org/10.1111/1468-2427.12416.
30. Shin, D., Aliaga, D., Tunçer, B., Arisona, S. M., Kim, S., Zünd, D., & Schmitt, G. (2015). Urban sensing: Using smartphones for transportation mode classification. *Computers, Environment and Urban Systems*, *53*, 76–86 https://doi.org/10.1016/j.compenvurbsys.2014.07.011.
31. Stenneth, L., Wolfson, O., Yu, P. S., & Xu, B. (2011). *'Transportation mode detection using Mobile phones and GIS information' no July 2011*.
32. Stopher, P., Clifford, E., Zhang, J., & Fitzgerald, C. (2008). *Deducing mode and purpose from GPS data, working paper ITLS-WP-08-06, Institute of Transport and Logistic Studies, University of Sydney*.
33. Stopher, P., Jiang, Q., & Fitzgerald, C. (2005). *Processing GPS data from travel surveys*, (pp. 1–17).
34. Strauss, J., Miranda-Moreno, L. F., & Morency, P. (2015). Mapping cyclist activity and injury risk in a network combining smartphone GPS data and bicycle counts. *Accident; Analysis and Prevention*, *83*, 132–142 https://doi.org/10.1016/j.aap.2015.07.014.
35. Ton, D., Cats, O., Duives, D., & Hoogendoorn, S. (2017). How do people cycle in Amsterdam, Netherlands? *Transportation Research Record: Journal of the Transportation Research Board*, *2662*(November), 75–82 https://doi.org/10.3141/2662-09.
36. Tsui, S., & Shalaby, A. (2007). Enhanced system for link and mode identification for personal travel surveys based on global positioning systems. *Transportation Research Record: Journal of the Transportation Research Board*, *1972*(April), 38–45 https://doi.org/10.3141/1972-07.
37. van de Coevering, P., De Kruijf, J., & Bussche, D. (2014). *Policy renewal and innovation by means of tracking technology Een innovatieve schakel tussen onderzoek en fietsbeleid*. Bijdrage aan het Colloquium Vervoersplanologisch Speurwerk 20 en 21 november 2014, Eindhoven. https://www.cvs-congres.nl/cvspdfdocs_2014/cvs14_033.pdf.
38. Wolf, J., Guensler, R., & Bachman, W. (2001). Elimination of the travel diary: Experiment to derive trip purpose from global positioning system data. Transportation research record. *Journal of the Transportation Research Board*, *1768*, 125–134.
39. Wolf, J., Schönfelder, S., Samaga, U., Oliveira, M., & Axhausen, K. W. (2004). Eighty Weeks of Global Positioning System Traces: Approaches to Enriching Trip Information. *Transportation Research Record: Journal of the Transportation Research Board*, *1870*(1), 46–54.
40. Xiao, G., Juan, Z., & Gao, J. (2015a). *Travel mode detection based on neural networks and particle swarm optimization*, (pp. 522–535).
41. Yang, F., Yao, Z., & Jin, P. J. (2016). GPS and acceleration data in multimode trip data recognition based on wavelet transform Modulus maximum algorithm. *Transportation Research Record: Journal of the Transportation Research Board*, *2526*, 90–98.
42. Zhang, L., Dalyot, S., Eggert, D., & Sester, M. (2012). *Multi-stage approach to travel-mode segmentation and classification of Gps traces*, *ISPRS - International Archives Photogrammetry Remote Sensing and Spatial Information Sciences, XXXVIII-4/(August)* (pp. 87–93) https://doi.org/10.5194/isprsarchives-xxxviii-4-w25-87-2011.
43. Zong, F., Bai, Y., Wang, X., Yuan, Y., & He, Y. (2015). Identifying travel mode with GPS data using support vector machines and genetic algorithm. *Information*, *2015*(6), 212–227 https://doi.org/10.3390/info6020212.

Publisher's Note