Transportation mode-based segmentation and classification of movement trajectories

Filip Biljecki*
Section GIS Technology, Delft University of Technology, The Netherlands. f.biljecki@tudelft.nl
Business development, Geofoto d.o.o., Zagreb, Croatia.
Hugo Ledoux
Section GIS Technology, Delft University of Technology, The Netherlands. h.ledoux@tudelft.nl
Peter van Oosterom
Section GIS Technology, Delft University of Technology, The Netherlands. p.j.m.vanOosterom@tudelft.nl

Cite as:
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), pp. 385-407.
DOI: http://dx.doi.org/10.1080/13658816.2012.692791

*Corresponding author

Abstract

The knowledge of the transportation mode used by humans (e.g.bicycle, on foot, car, and train) is critical for travel behaviour research, transport planning and traffic management. Nowadays, new technologies such as the GPS have replaced traditional survey methods (paper diaries, telephone) since they are more accurate and problems such as underreporting are avoided. However, although the movement data collected (timestamped positions in digital form) have generally high accuracy, they do not contain the transportation mode. We present in this paper a new method for segmenting movement data into single-mode segments and to classify them according to the transportation mode used. Our fully automatic method differs from previous attempts for five reasons: (1) it relies on fuzzy concepts found in expert systems, i.e.membership functions and certainty factors; (2) it uses OpenStreetMap data to help the segmentation and classification process; (3) we can distinguish between 10 transportation modes (incl. between tram, bus, and car) and we propose a hierarchy; (4) it handles data with signal shortages and noise, and other real-life situations; (5) in our implementation, there is a separation between the reasoning and the knowledge, so that users can easily modify the parameters used and add new transportation modes. We have implemented the method and tested it with a 17-million point dataset collected in the Netherlands and elsewhere in Europe. The accuracy of the classification with the developed prototype, determined with the comparison of the classified results with the reference data derived from manual classification, is 91.6 percent.

Keywords: Movement trajectory, GPS track, travel behaviour research, OpenStreetMap

1 Introduction

The knowledge of the transportation mode used by humans (e.g.bicycle, on foot, car, and train) is critical for applications such as travel behaviour research (Bohte and Maat, 2009) where researchers aim at understanding human travel behaviour in order to predict travel patterns and evaluate transport-related measures and policies. Travel behaviour is concerned with how people travel, where they go, how often, which transportation mode do they use, whether they chain trips, which route they choose, and so on. Researchers try to understand the impact that the built environment, the quality of the public transport and the cost of various transportation modes have on humans. This knowledge can also be used for transport planning and traffic management, see for instance Asakura *et al.* (2000) or Ranjitkar *et al.* (2002).

In the past, the data required by travel behaviour researchers were usually acquired in travel surveys, involving randomly sampled individuals. Researchers collected the information of the transportation mode used through paper diaries filled by participants or telephone surveys, which often resulted in underreporting of short trips and in inaccurate and incomplete data (McGowen and McNally, 2007). Recent advancements in positioning technologies—such as the Global Positioning System (GPS)—have enabled inexpensive and straightforward acquisition of movement data, but they come in a different form: sequential timestamped positions: (x_1, y_1, z_1, t_1), (x_2, y_2, z_2, t_2),

$\ldots, (x_n, y_n, z_n, t_n)$. The advantages are many: underreporting of trips is less likely, the data are immediately available in a digital form and can be analysed in a geographical information system (GIS) environment, and in general more data are available at a finer level of resolution (Bricka and Bhat, 2006; Wolf, 2000; Draijer *et al.*, 2000). Further, most researchers conclude that these receivers have now completely replaced, rather than supplement, traditional travel diaries. It should be noted that several travel surveys with positioning loggers have already been done, see, among others, Draijer *et al.* (2000); Bohte and Maat (2008); Axhausen *et al.* (2004).

However, in contrast to travel diaries and surveys, these techniques do not collect the transportation mode. Combining the use of receivers with traditional paper/telephone surveys would be a high burden for participants of these surveys (Wolf *et al.*, 2001), and since the datasets are usually vast, manual classification may not be possible.

We present in this paper a new method to automatically detect and classify a movement trajectory (such as a GPS log) for the transportation mode. Since a trajectory may contain multiple transportation modes, the problem is extended to the segmentation of the movement data into single transportation modes; we introduce our terminology in Section 2. As explained in Section 4, the segmentation works by detecting potential transition points between two transportation modes at brief stops at train stations, traffic lights, bus stops, etc. Each segment between consecutive potential transition points is classified, and adjacent segments with the same classification outcome are merged in an iterative process. For each trajectory, various numerical values (we call them indicators), which contribute to the identification of the transportation mode, are calculated. Some of these indicators are derived from the geographic data (e.g.the proximity of the trajectory to the tram network). As explained in Section 5.1, we use the geodata from OpenStreetMap, which is free to use and of good quality (at least in Western Europe). The transportation modes are classified by analysing the indicators with an explicit knowledge base set with a number of empirically derived fuzzy membership functions; our method relies thus on fuzzy concepts found in expert systems, i.e.membership functions and certainty factors. Finally, the classification results have a certainty value. The classification of data gaps (e.g.caused by a signal shortage during the logging of a trajectory, which often arise with our test datasets) is also addressed.

The method we propose has two main advantages over previous work: (1) a more extensive and detailed list of transport modes is used, we differentiate between 10 modes while previous work was often limited to 4 or 5; (2) it tries to handle errors in tracks (due to signal shortage for instance) and can thus be used both with older devices and new ones. A novelty of our method is that we introduce a hierarchy of transportation modes, and for each segment to be classified we assign the mode only if we are sure, if not we return a transportation mode lower in the hierarchy.

We have implemented our method and we have tested it with, among others, a 16-million point dataset that was collected in 2007 in the Netherlands for a study (Bohte and Maat, 2009), which contains many real-life cases useful for checking the robustness of the method. We report in Section 6 on this experiment, and we discuss the results we obtained against a semi-manual classification that had been performed during the study. At this moment, our prototype permits us to classify movement trajectories for 10 transportation modes, but since there is a clear separation between the reasoning engine and the knowledge, it can easily be extended by users so that new transportation modes are considered.

2 Terminology

Moving objects are all objects that may change their position through time (e.g.people). In this case their position can be often represented with a point, without losing valuable information. During their existence, moving objects experience *journeys*, each one occupying a time interval in the object's lifespan and moving the object between two relevant locations—bird migration, daily commuting, and mail service. Any movement, including journeys, can be perceived as countable traveling units—"a record of the evolution of the position of an object that is moving in space during a given time interval in order to achieve a given goal." (Spaccapietra *et al.*, 2008).

The movement of an object may be *segmented* into trajectories between two relevant locations. This segmentation is application-dependent. For example, the movement of a truck of a delivery company can be segmented into daily movements, but also into movements between customers.

In this paper, two varieties of segmentations are considered. First, the segmentation into separate *journeys*, which we define as connections between two relevant locations related to an individual or a household, e.g.the movement from home to work and from work to shopping (Maat and Timmermans, 2006). Second, since trajectories can be undertaken with the use of different transportation modes, another segmentation is established for obtaining single-mode trajectories, simply denominated as *segments*. In the segmentation of the trajectories for discerning different transportation modes, the points where the segmentation occurs are defined as *transition points*.

The record of a movement is synonymous with a *track*, which is more applicable in the context of current acquisition technologies. The recording is nowadays generally done by sampling (observing) positions in a certain interval of time, deriving *sampled points*—sequences of positions and timestamps (i.e.position in space-time) in a specific time interval.

In order to formalise the presented concepts and related terms with their relations, a UML class diagram, inspired by the work of Verbree *et al.* (2005), is given in Figure 1.

A sampled point is part of a single transportation-mode segment. Each point consists of a timestamped position, but with additional information it is possible to derive supplementary information; for instance it is possible to calculate its speed from the distance and time difference to the subsequent point.

The first and last point of a segment are transition points, which separate the segment from adjacent segments completed with other transportation modes. A segment is part of a journey, another collection of points, but related to a purpose of movement between two relevant locations—e.g.commuting. A relevant location, the point separating two consecutive journeys, can also have a type (e.g.home, shop).

A movement archive contains all journeys of an individual in a recorded timeframe.

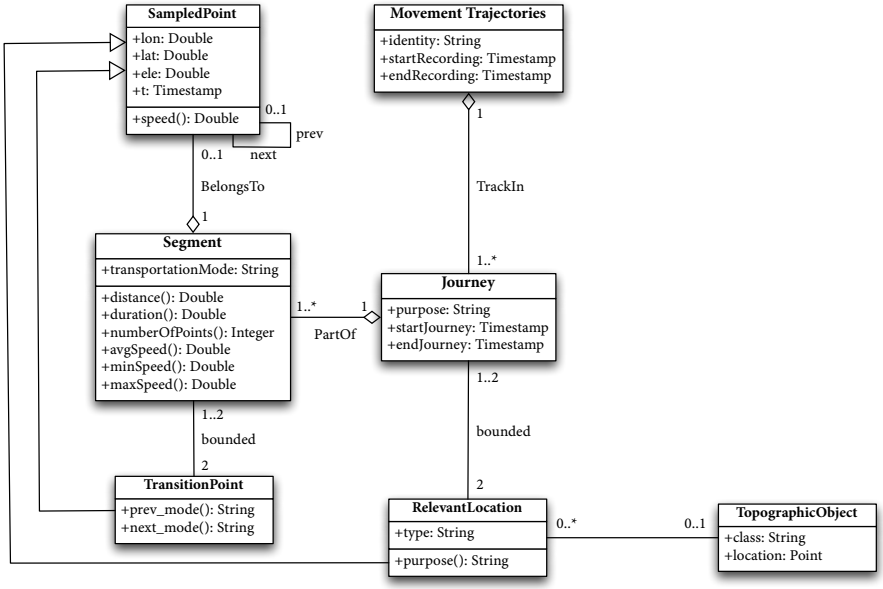


Figure 1: UML class diagram formalising the presented concepts relevant to segmentation and classification of movement trajectories. Adapted from (Verbree *et al.*, 2005).

3 Related work

There are several publications describing attempts to solve the problem presented in this paper. Most publications concentrate on the classification and omit the segmentation problem (Byon *et al.*, 2009; Dodge *et al.*, 2009; Reddy *et al.*, 2008). On average, the published methods classify between four and five transportation modes and use around four indicators. The accuracies of the studied methods are mostly between 70 and 85%. Table 1 gives an overview of the main methods, with their main characteristics.

In general, all methods use the speed between two consecutive points as the primary variable for mode detection, implying that the speed gives the highest indication of a transportation mode (Bohte *et al.*, 2008; Schüssler and Axhausen, 2009). Because different transportation modes have similar speeds (e.g.cars, trams and trains), additional knowledge is essential in order to distinguish modes. Apart from the average speed, a few methods use the maximum speed in a trajectory as an additional indicator from the knowledge of the speeds at each observation (Stopher *et al.*, 2008). Researchers note that nearly maximum values should be used rather than maximum values of speeds and acceleration in order to make the method robust for noisy measurements (Stopher *et al.*, 2007; Schüssler and Axhausen, 2009). While there is not a strictly defined value, nearly maximum values are usually calculated with 95th or 85th percentiles.

Geodata is not frequently used for calculating the indicators or facilitating the segmentation and classification, but methods using geodata report higher accuracy (up to 95%) (Gonzalez *et al.*, 2010).

Table 1: Comparison of the reviewed methods for transportation mode identification. (The dash represents unknown information.)

Method	Modes	Criteria	GIS data usage	Accuracy (%)
(Byon *et al.*, 2009)	4	3	no	82
(Schüssler and Axhausen, 2009)	5	3	no	—
(Zheng *et al.*, 2010)	4	5	no	75
(Bohte *et al.*, 2008)	4	2	yes	70
(De Boer, 2008)	7	6	yes	—
(Dodge *et al.*, 2009)	4	3	no	82
(Reddy *et al.*, 2010)	4	3	no	74
(Liao *et al.*, 2007)	3	2	yes	—
(Gonzalez *et al.*, 2010)	3	8	no	91
(Lester *et al.*, 2008)	4	3	yes	—
(Stopher *et al.*, 2008)	7	4	yes	95
Average	4.5	3.8	5 of 11	81.3

Geodata may be used not only for detecting line infrastructure features (e.g.roads and railways), but also for determining potential transition points such as railway stations (Liao *et al.*, 2006). In addition, underground modes (metro) can be detected by finding signal shortages with last known points around the locations of the stations (Shalaby *et al.*, 2006; Stopher *et al.*, 2008).

One major problem of related methods is that they do not segment a trajectory into single-mode segments. Assuming the use of a single transportation mode may result in a wrong classification since people often use multiple transportation modes while travelling. Zheng *et al.* (2008a) highlight that fact, stating that a person usually walk between the use of 2 transportation modes. In our method, we exploit that fact: in order to detect a transition, first we try to find walking segments.

Furthermore, several researchers do not address problems with data such as occasional gaps caused by signal shortages and noise. During our experiments, errors and noise were very frequent (see Section 6), especially in data acquired with older GPS receivers.

Most methods consider only a limited number of transportation modes, which may be trivially distinguishable in most circumstances due to their very different behaviour in movement. Methods which incorporate more transportation modes usually do not report high accuracy—a negative correlation between the number of modes and accuracy can be observed—and the methods usually derive single results without a value of certainty, with no alternative result.

From the classification perspective the most common approach is the use of a decision tree-based method, which delivers single results without a value of certainty and does not consider ambiguity when two modes have similar behaviour (Zheng *et al.*, 2010, 2008b; Bohte *et al.*, 2008; Bohte and Maat, 2008, 2009; Lester *et al.*, 2008).

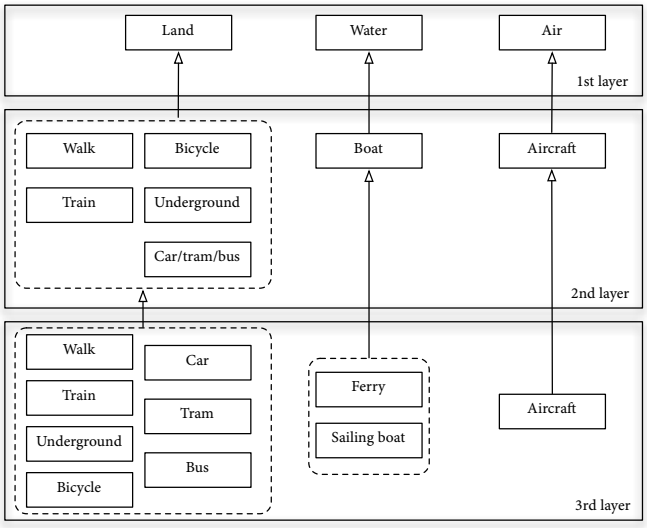


Figure 2: Hierarchy of transportation modes.

4 Our method

The method we propose first segments a movement trajectory into single-mode segments. The segments are then passed to the classification system, and in the post-classification phase, consecutive segments with the same transportation mode are merged (Figure 3).

Our method primarily relies on the use of a fuzzy expert system solution with indicators—numerical values that may indicate the use of a set of transportation modes. As explained in Section 4.3, the classification is empirically manually trained with training data (supervised learning), in which relationships are found from the calculated indicators to a particular transportation mode, and these are realised with membership functions.

We consider the most frequent transportation modes in Europe. Their list is composed from the recent Dutch National Travel Survey (Ministerie van Verkeer en Waterstaat, 2009), with the addition of sea and air transportation modes, and metro (underground). The complete list of the modes is shown in Figure 2. Notice that the inclusion of the sea and air transportation modes is novel in relation to the existing solutions.

For technical reasons, we introduce a new class "Stationary" for all non-moving points. It is not shown in the Figure 2 as it is auxiliary.

As one may anticipate, in a few cases, discerning between a certain subset of the listed modes may not be possible with a high certainty. For example, buses and cars have similar speeds and acceleration in urban areas and both operate on the same infrastructure. We therefore introduce a hierarchy of transportation modes in order to give an accurate result, which is more acceptable

than returning inaccurate or uncertain results. Three layers of transportation modes are generated, and the classification is done separately for each layer.

The first basic layer contains the most general groups: land, sea, and air. In some cases it may be complex to distinguishing the following groups of modes:

- Bus, tram, and car (similar speed, and use of the same infrastructures);
- Sailing boat and ferry.

In order to avoid the possible errors of the classification system, the second layer contains aggregations of some *ambiguous* modes. Hence, the second layer comprise seven transportation modes: walk, bicycle, car/tram/bus (a single *mode*), train, metro, boat (comprises sailing boat and ferry), and aircraft. The third layer has the car, tram and bus modes, and the sailboat and ferry models, as separate modes.

4.1 Selection of the indicators

Our research involved testing the usability of a large number of numerical values derived from the timestamped positions for the classification of the trajectories. The selection of the indicators suitable for the classification for transportation modes resulted in nine values:

- three single values of speeds in the segment: its 95th percentile (the nearly maximum speed), the mean speed, and the mean moving speed,
- five average proximities of the segment to the infrastructures used by the selected transportation modes (railway, tram lines, roads, bus lines, metro lines—with segments that are not underground where might be GPS reception), and
- the location of the trajectory with respect to water surfaces.

According to other researchers, the acceleration is as a useful indicator (Zheng *et al.*, 2010). However, experiments with our datasets have shown otherwise. Despite the speeds in most of the GPS devices being measured accurately with the Doppler effect (Zhang *et al.*, 2006), because of insufficient sampling periods (equal or above five seconds) and because of the variation of speeds between the samples, the acceleration was not accurate enough to be used as indicators†. Furthermore, in our experience, when considering a larger number of transportation modes the acceleration is no longer an indicator that helps the segmentation process.

On a related note, since the average moving speed is computed from consecutive positions which may contain positioning errors, a stationary GPS device will often record low speeds when not moving at all. This issue is taken into account, it is detected with an algorithm we developed and such data is filtered out.

†With the use of newer devices with a sampling rate of 1s, acceleration could however be used.

4.2 Concept of the segmentation

As described in Section 2, a GPS track may have been completed with multiple transportation modes, therefore before any classification first we need to divide it into single transportation-mode segments. The segmentation is done in a two-step process:

1. partition of trajectories to single-journey segments (between two meaningful locations), and
2. segmentation of journeys into single-mode segments.

Although both segmentations technically derive segments, the segments in the first segmentation are referred to as *journeys*, and the latter simply as *segments*, as visible in the UML diagram in Figure 1. Once a trajectory is segmented, it is ready for its classification. Therefore, the trajectories are segmented before any knowledge of the transportation mode.

4.2.1 Segmentation into journeys between two relevant locations

Different journeys are often separated by a longer interruption in logging the data, caused by either a signal shortage (individual in a building) or a device turned off. However, regular signal shortage while travelling (e.g.entering a tunnel, or a journey with train) often exhibits the same behaviour. As a consequence, in addition to the time difference, the distance between the last known point before the gap and the first point after the interruption of logging is taken into account. In case of journeys, and not of a signal shortage while moving, the departing point of the next journey is usually close to the arrival point of the previous journey. By examining several datasets we concluded that most of the journeys are mutually separated by longer period such as a working shift (8-9 hours) or a night, hence they are straightforward to detect and segment.

4.2.2 Segmentation into single-mode segments

The second step of the segmentation is more challenging: the transitions between modes occur much faster than transitions between journeys and they require a different approach.

Mountain and Raper (2001) report that the a rapid and sustained change in direction or speed indicates a change of mode. Therefore, a segmentation algorithm would require detecting sudden changes. Although theoretically such approach is plausible, a serious problem arises when dealing with a transition which does not bear a noticeable change in speed and/or direction.

Liao *et al.* (2006) segment multi-modal trajectories by analysing the proximity to potential-transition locations such as bus stops. This method presents another interesting use of geodata. However, their approach may have difficulties in areas with dense traffic features (especially in the Netherlands), where the distance between potential transition points for various modes may be in the range of GPS errors, hence this method is used only partially in order to discern between cars, buses and trams (we elaborate on this in Section 4.4).

Zheng *et al.* (2010) indicate that a person usually walks or stops during the transition. By examining the test dataset and observing the same behaviour, we choose to second their conclusions and to follow this logic. However, by examining the available data we have noticed that the transitions often cause data interruption (signal shortage under the roof in a train station, or entering a bus), hence signal shortages have to be added to the list. They are used as an additional indication for a potential transition.

All stops which are longer than a specific threshold, and also those before a signal shortage, are considered as *potential transition points*. These events indicate that the transportation mode *might* have changed. In determining the threshold, one should consider that over-segmentation of the trajectories is better than under-segmentation since fast transitions may pass undetected (e.g.exiting a tram/bus, and immediate departure with some other modes).

Since many single-mode trajectories contain stops (e.g.cars stopping for traffic lights), initially the trajectory may be segmented into a high number of segments. This is however not a problem since we merge as post-processing consecutive segments having the same mode (see Figure 3).

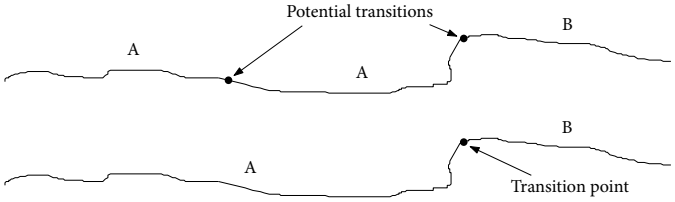


Figure 3: All stops and signal shortages in a trajectory are first marked as points of potential transition, and the segments are classified separately. If two adjacent segments have the same classification outcome, they are merged into one segment.

Each segment is terminated after a stop or signal shortage is encountered. The threshold for a data disruption is set to 30 seconds, while a stop is considered when there is no movement for more than 12 seconds. Since in a stop the position might not be recorded at exactly the same position and there might be slight movement, a stop is detected when consecutive points in an interval of 12 seconds do not have a speed higher than 2 km/h.

Despite our efforts, some cases of very fast transitions cannot be detected, one example being when a person is running to a tram which immediately departs. These cases would require a different approach since a shorter threshold would compromise other results by creating too many segments with just a few points. However, in practice such cases are not frequent.

4.3 Concept of the classification

There are various general approaches for building a classification system, as it can be concluded from the different existing approaches described in Section 3. In this paper, an expert system approach is chosen because of its maturity and because it has been used successfully at solving other problems (Holzmann *et al.*, 1999; Rearden *et al.*, 2007; Wentz *et al.*, 2008).

An expert system is a software package that can reason through complex situations. It comprises the knowledge of an expert in a certain field to provide answers to problems (Buchanan and Duda, 1982). It is applicable to specific problems and has been developed to substitute experts. The most typical usage of expert systems is in medicine (Grazia, 2006). Expert systems have been used in GIS, for instance in cartography (Van Oosterom *et al.*, 2001; Alkemade, 2000; Kotte, 2002), and for area/object classification of topologically structured topographic data converted from spaghetti data (Van Oosterom, 1999).

Fundamentally, expert systems consist of a knowledge base (evidences *e*), and an inference procedure (rules), which derive conclusions (hypotheses *h*): IF *e* THEN *h*.

Another important concept in expert systems is (un)certainty, which occurs when one is not absolutely certain about a piece of information (Nickles and Sottara, 2009). The degree of certainty, introduced by Shortliffe *et al.* (1975), is represented by a numerical value CF(*h*, *e*), where CF is the certainty factor, a quantification of the confidence that an expert might have in a conclusion or hypothesis *h* that s/he has arrived at from an evidence *e*.

In case that a set of conclusions (hypotheses) derives multiple values of CFs, they should be propagated through a reasoning chain, i.e.combined, to obtain one single certainty factor. Several inference methods had been established for this operation. For instance, in MYCIN (an early expert system developed in the early 1970s at Stanford University), when two CFs are ANDed (conjunctive reasoning), the joint CF is the minimum value of the two (Shortliffe and Buchanan, 1975):

$$\mathrm{CF}[A \cap B] = \min(\mathrm{CF}[\mathrm{A}], \mathrm{CF}[\mathrm{B}])$$

This approach has the advantage that one CF with the value of 0 may result in a joint CF of 0, i.e.if there is strong evidence that contradicts a hypothesis, other hypotheses with a non-zero CF are discarded.

The presented concepts appear to be suitable for solving the problem of the classification of movement trajectories.

In relation to this project, an example of a fact is the mean speed of a trajectory—30 km/h. By considering solely the mean speed of a trajectory, there is suggestive evidence that the value of the speed *probably* represents a car, with e.g.a CF of 1.0.

The fuzzy expert system developed for this paper uses fuzzy logic to derive certainty factors, i.e.*fuzzy variables are used to assign certainties to each derived hypothesis*. Consider the following case of a rule as an explanation of the concept. If the maximum speed in a segment is 118 km/h, from common sense we can build a rule that concludes that the transportation mode could be a car: $\mathrm{CF}_{\mathrm{car}}^{\mathrm{max.speed}}(118\ \mathrm{km/h}) = 1.0$.

While higher speeds for cars are rare, they should not be discarded as a possibility, since there *might* be a possibility that the segment was completed with a car. In order to retain the reasoning, but give it less weight, this is done for instance by assigning a lower CF: $\mathrm{CF}_{\mathrm{car}}^{\mathrm{max.speed}}(138\ \mathrm{km/h}) = 0.6$.

Therefore, the certainty factors in fuzzy expert systems are a function of available evidence: CF = $f(e)$, i.e.membership functions which are empirically defined by investigating travel behaviour for each transportation mode in the training data, a subset of the test data used for that purpose.

Each available fact should be used for each considered transportation mode (class) in the system. For instance, extending the use of the information of the maximum speed for trains: $\mathrm{CF}_{\mathrm{train}}^{\mathrm{max.speed}}(118\ \mathrm{km/h}) = 0.4$.

Therefore, each rule in the system determines an array of certainty factors, one for each mode considered:

IF (max. speed is 118 km/h)
THEN $\mathrm{CF}_{\mathrm{car}}^{\mathrm{max.speed}} = 1.0$, $\mathrm{CF}_{\mathrm{train}}^{\mathrm{max.speed}} = 0.4, \ldots$

In case of multiple facts, the final CF is determined as a conjunctive CF since the rules are not used in a particular sequence:

IF (max. speed is 55 km/h)
THEN $\mathrm{CF}_{\mathrm{tram}}^{\mathrm{max.speed}} = 0.85$

IF (average proximity to tram network is 4933 m)
THEN $\mathrm{CF}_{\mathrm{tram}}^{\mathrm{prox.}} = 0$

$\rightarrow \mathrm{CF}_{\mathrm{tram}} = \min(0.85, 0) = 0$

This is done for each mode. From the last example, it is visible that one rule in such system could completely eliminate the possibility of a transportation mode based on only one fact. Therefore, the presented classification system works on the *elimination of unlikely modes* by assigning them CFs of zero for each evidence that is strongly against a hypothesis.

In order to formalise the presented concepts an overview is given. For each transportation mode m (e.g.train) of the N considered modes (modes $m_1 \ldots m_n$), the classification system contains k membership functions f_m^i, where k is total number of indicators (facts) used as the input of the classification and i marks the designation of the indicator, e.g.f_2^3 or $f_{\mathrm{train}}^{\mathrm{max.speed}}$. For each segment, k indicators $i_1 \ldots i_k$ are calculated (e.g.i_3 or $i_{\mathrm{avg.speed}}$) and passed to the respective membership functions for each transportation mode (e.g.$f_{\mathrm{train}}^3(i_3), f_{\mathrm{car}}^3(i_3), f_{\mathrm{bicycle}}^3(i_3), \ldots$) from which certainty factors $\mathrm{CF}_m^i = f_m^i(i)$ are calculated. The total number of the membership functions and corresponding certainty factors is the product of the number of indicators k with the number of the considered transportation modes n.

After computing the k certainty factors for each transportation mode, the system determines the minimum value for each and considers it as a the final CF. The confidence that the mode in question was used to complete the classified segment is:

$$
\begin{array}{llll}
\mathrm{CF}_1^1 = f_1^1(i_1) & \mathrm{CF}_1^2 = f_1^2(i_2) & \ldots & \mathrm{CF}_1^k = f_1^k(i_k) \quad \Rightarrow \mathrm{CF}_1 = \min(\mathrm{CF}_1^1, \ldots \mathrm{CF}_1^k) \\
\mathrm{CF}_2^1 = f_2^1(i_1) & \mathrm{CF}_2^2 = f_2^2(i_2) & \ldots & \mathrm{CF}_2^k = f_2^k(i_k) \quad \Rightarrow \mathrm{CF}_2 = \min(\mathrm{CF}_2^1, \ldots \mathrm{CF}_2^k) \\
\vdots & \vdots & \vdots & \vdots \\
\mathrm{CF}_n^1 = f_n^1(i_1) & \mathrm{CF}_n^2 = f_n^2(i_2) & \ldots & \mathrm{CF}_n^k = f_n^k(i_k) \quad \Rightarrow \mathrm{CF}_n = \min(\mathrm{CF}_n^1, \ldots \mathrm{CF}_n^k)
\end{array}
$$

The mode with the highest (non-zero) CF may be considered as the result of the classification.

A significant advantage of using certainty values in the results is the possibility of sorting the results by the value of CF and obtaining alternative results in order to improve the performance of the classification system.

The knowledge used for the classification is stored (encoded) separately in membership functions, and it is explicitly defined. There are numerous types of membership functions. A common construction of a membership is trapezoidal, and it is used in our implementation. Figure 4 shows the set of membership functions of the considered modes for the indicator of the nearly maximum speed.

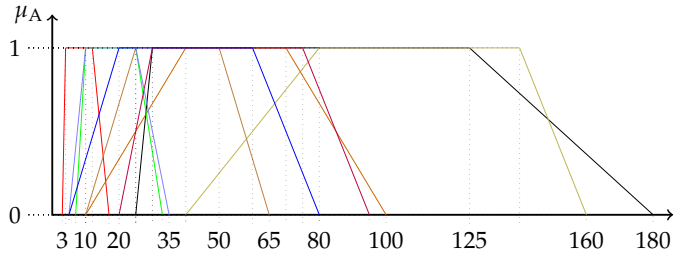


Figure 4: The membership functions usually overlap. This is an example for the membership functions for nine modes used in the indicator of the nearly maximum speed (in km/h). The following modes are plotted: car (black), train (dark yellow), walk (red), bicycle (green), tram (brown), bus (purple), sailing (light blue), ferry (blue), and underground (dark orange). The classes stationary and aircraft are left out for aesthetic reasons.

This is also one of the simplest constructions, and it is suitable for this approach. It requires the definition of four points, where $\leq x_0$ and $\geq x_3$ correspond to a certainty of zero, while between x_1 and x_2 to one. Every value in between the four points is considered as fuzzy. It is important to note that in this concept the range of the derived values by the MF is [0,1].

The definition of the membership function for each indicator for each transportation mode is done in a training process also known as trial and error (Section 5.3).

4.4 Discerning between similar modes

Classification between bus, car and tram is usually ambiguous because of comparable speeds and because they both use the road network.e While buses and cars in urban areas operate on the same

roads, when a segment is detected far from the bus network the classification rejects bus as a possibility by assigning a CF of zero. However, in case of the presence of a bus line, but in some cases also a tram line which operates adjacently to the road (on sometimes on the road) and it is in range of GPS errors, the classification is ambiguous and the three classes are assigned with the comparable CF, e.g.1.0. This situation is solved by using the locations of the bus and tram stops, and by using the knowledge of the previous used mode. Indeed, if the segment started at a bus/tram station then there is a high probability that the segment was completed by a bus or tram. Thus, the certainty factors for these modes are increased; Figure 5 clarifies this theory. If the new segment is started close to a station, then the corresponding mode gets a CF increase in the subsequent segment. The value is currently put to 0.2 since we have noticed that virtually all discrepancies between these three modes in the classification are less than 0.2. The size of the buffer iss currently set to 20 m which compensates the size of the bus/tram stops and the GPS noise.

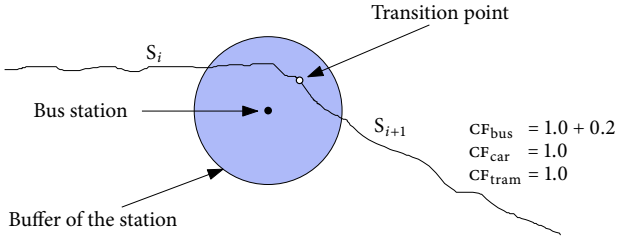


Figure 5: Injecting certainty factors supplement for segments which commence at a station for bus or tram contribute to the distinction of the modes car, bus, and tram.

However, many tram and bus stops are close to regular car stops (such as traffic lights), which may wrongly assign a car segment to a bus or a tram if it was started close to a station. Hence, our method takes into account the knowledge of the previously transportation mode: if the previous segment was completed with a car, and ambiguity between car, tram and bus exist, the class car gets a favourable CF supplement. Bus and tram segments may be possible only when the previously used mode was walking.

In addition, buses and trams may stop at points outside the buffers of stations (e.g.bridges), which may cause the opposite classification. The knowledge of the previously transportation mode is then taken into account as well.

The disadvantage of this approach is that in rare situations where a person was dropped off from a car directly at a bus or tram station and continued the journey with either a bus or tram are from that point wrongly classified as car segments. This is partially solved by analysing the dwell time between two segments. If the dwell time after a non-walking segment is longer than a certain threshold and took place in the buffer of a station, then it is assumed that the person was waiting for a bus/tram, rather than waiting in a car for a traffic light.

4.5 Dealing with disruptions in the data

Signal shortages that cause disruptions in the acquisition of data (i.e.gaps) are frequent and hard to handle since we are dealing with the classification of *non-existing* data. As noted, we consider data as missing when no samples are recorded for more than 30 s. The problem is complex since there are numerous cases, one example is that the transportation mode could have changed during the disruption.

Resolving the gaps requires investigating many possible cases that occur in practice. In addition to these problems, this method takes advantage of gaps, since the metro mode does not have any reception, and it is detected by the disruption of signal in between entrances to the two metro stations, similar to the methods of Stopher *et al.* (2008) and Shalaby *et al.* (2006).

The following distinct cases account for most, if not all occurrences of gaps, and their reconstruction was developed and implemented in the prototype. All cases are depicted in Figure 6.

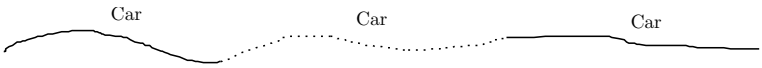
Since the points are not sampled, we must *guess* what happened during the gap. The distance between the two adjacent recorded segments is known, along with the time difference. From these, the average distance may be computed, although this is rather a rough approximation due to the potential sinuosity of the travelled path. As one might suggest, proximity to the stations for certain modes are available for the points on the edge of the gap. Although it is possible to take into account the proximity to the stations, these cases have something more in common—they occur on the network of each corresponding mode. Hence, instead of stops, the location of networks is used, which are already available from the preprocessing procedure.

Before each disruption in the data, the system stores the classification result of the preceding segment, and the distance from the last known point to all considered infrastructures. This is also done for the first point after the gap.

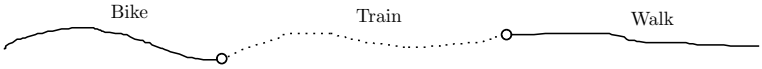
The reasoning system analyses the infrastructures from the buffers of the boundary points of the segments (last recorded point in the previous segment and first in the subsequent segment). If two infrastructures match (e.g.if both points fall in a buffer), then the corresponding mode is assigned. This is especially useful for underground modes since it is the only way to classify them. Indeed, some metros have part of their networks on the surface, but almost always involve data missing time intervals.

In case of the match of multiple infrastructures, the average speed of the gap and the knowledge of the previous transportation mode prevails. If neither conditions are met, the system analyses the average speed of the gap, and the average speed of the first 20 points of the next segment. If either speed is higher than 300 km/h, the gap is marked as 'air'.

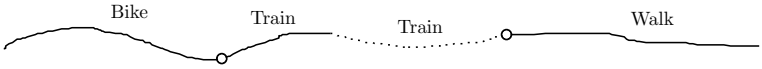
The value of 20 points in the following segment was taken into account in order to preserve the travel behaviour of the segment in the portion closer to the gap. This is useful in cases where the subsequent segment is relatively long and during its course exhibits behaviour which can be different than in its part closer to the previous segment (e.g.much higher speed).



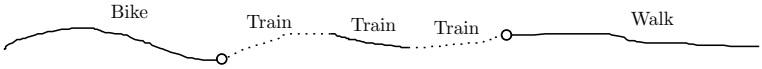
(a) ”Regular” gaps where the mode was not changed.



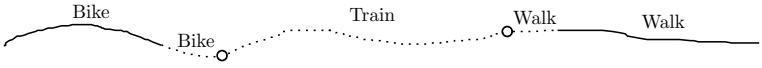
(b) The whole segment done with another mode is not recorded.



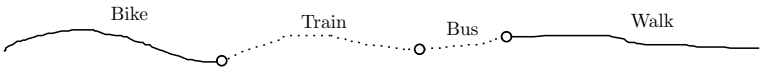
(c) The first transition is recorded, but not the rest of the segment and the second transition.



(d) Neither of the transitions is detected, but a small fragment of the segment is recorded in between.



(e) Signal shortage occurs during both transitions, the mode is not recorded, and the disruptions exist in the earlier and next segments.



(f) Signal shortage occurs during both transitions, the mode is not recorded, and the disruptions exist in the earlier and next segments.

Figure 6: General cases of data interruption.

This solves the cases (a), (b), (c), (d) shown in Figure 6 by a unified approach. The method is also useful in segments where only fragments of data are available.

Case (e) is resolved only in instances where the time difference and distance to the occurred transition is small. In other cases the segment is marked as unknown as it involves too much ambiguity. The same applies for (f) which is a case that cannot be solved even with human intervention.

Another specific case that could not be solved is that if a person lost GPS signal while boarding a ferry, and reappeared after the segment was finished. The sea mode could not be resolved since both boundary points fall onto land. Reasoning that the person crossed a water polygon in between requires complex GIS operations (and there is always the possibility that the person crossed a bridge).

Although machine reasoning in signal shortages is complex, we have obtained satisfactory results and ”repaired” the available datasets. We believe that the presented method of classifying deficient and broken data is a contribution in this field.

5 Implementation

A prototype, implemented in Python, was created in order to test the presented approach. Figure 7 depicts the workflow of the implementation.

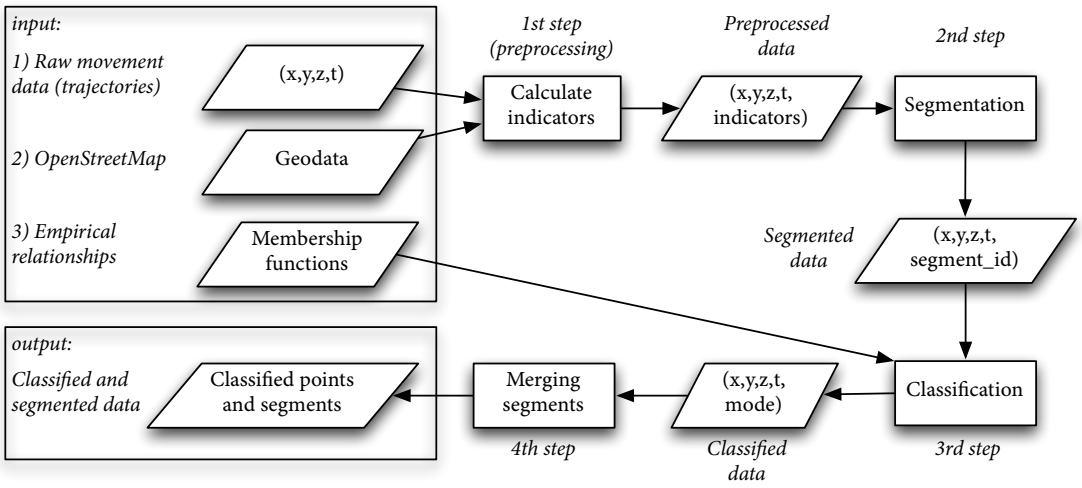


Figure 7: Flowchart of the implementation of the prototype.

Raw movement datasets in form of timestamped positions (x, y, z, t) are imported in a database (PostgreSQL with PostGIS in our case), and are preprocessed for the required indicators (first step). Geodata required for the calculation of the indicators come from the OpenStreetMap project (described in Section 5.1) and have been stored in the same database. To each point is attached a series of indicators which are later used in the classification. Afterwards the data is segmented (second

step) into single-mode trajectories in single-journeys (Section 4.2), which are then passed to the classification system. The classification system, aided by the membership functions and certainty factors (Section 4.3), classifies each point for the transportation mode (third step). The classified segments are then finally merged with adjacent segments of the same class (fourth step) resulting in the segmented and classified trajectories.

In order to test the developed method and the prototype two large movement datasets were used:

- the data from the survey conducted in the Netherlands by Bohte and Maat (2009) as part of a travel behaviour study focused on residential choice. This dataset contains 7-day movement logs of a thousand respondents collected with a handheld GPS logger with a SiRFStarII chip, with an average sampling period of 6.5 s. The data have been classified in an interpretation-validation process in which the system first made a preliminary basic segmentation and classification, after which the data have been checked and corrected by the respondent in a web-based questionnaire. Bohte and Maat (2009) are interested in removing or shorten the validation process by improving and automatizing the segmentation and classification process, which is one of the motives for our research. The data corrected by the respondents may be used as a reference for experiments and for validation.
- The manually classified data from the project of Van der Spek *et al.* (2009) from the Department of Urbanism, Faculty of Architecture, Delft University of Technology is used as well. The project concentrates on collecting data on various types of pedestrian movement in city centres. It addresses the topic of improving city centres for pedestrians, especially for shoppers and tourists (Van der Spek, 2010). This dataset has been collected with devices with a newer and more sensitive chip (SiRFStarIII), with a sampling period of 5 s.

Notice that for all the GPS tracks, we used the speed as calculated by the receivers (using the Doppler effect). Also, it was important to check the robustness of the method with data obtained with various devices of different sensitivity and technology (different frequency of signal shortages and accuracy). Thus, we downloaded additional movement data from the internet (e.g.OpenStreetMap raw logs for which the transportation mode was known), and from various devices that we used (mobile phones, and with handheld GPS devices of different production years and manufacturers).

In total, the available data contains 17.5 million GPS points in trajectories longer than half of a million kilometres, well covering the considered transportation modes and various movement scenarios required for testing the robustness of the method for different situations.

As an impression of the available datasets and their size, Figure 8 shows the Dutch city Amersfoort "mapped" from the available movement data. Frequently used paths (e.g.highways) can be observed by the aggregation of multiple points (i.e.thicker lines).

After the preprocessing (calculation of the indicators), the data are ready for segmentation, where the movement archive is segmented between all stops, and passed to the classification system.

It should also be said that the definition of the membership functions, *the empirical relationships*, are stored in a separate XML file which enables the extension of the prototype for additional indicators or transportation modes.

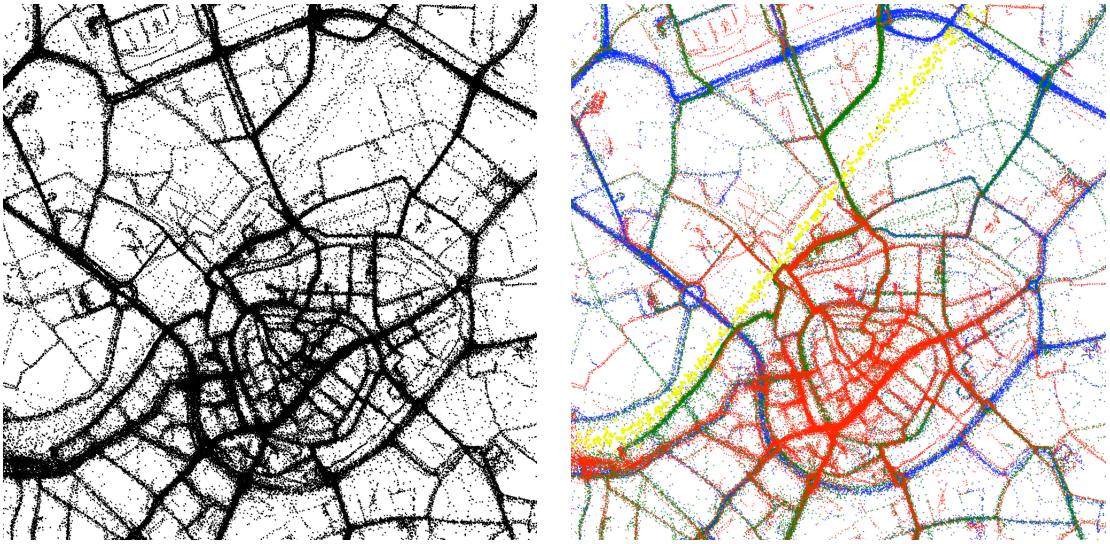


Figure 8: Visualisation of an excerpt of the GPS data used in the implementation (Amersfoort, the Netherlands) and its classification. The left side of the Figure represents a collection of raw GPS points, while the right side represents the visualisation of classified GPS data of the same spatial extent with different colours accounting for different transportation modes. Red represents walking, while green and blue are for bicycles and cars, respectively. Train is represented in yellow, visible on the diagonal railway.

5.1 OpenStreetMap

For geographical data, we chose to use the data from the OpenStreetMap (OSM) project because they are available for free usage, they contain all the needed features (e.g.bus infrastructure, at least in the Netherlands) and they have a good coverage (at least in Western Europe). OSM is a collaborative project to create a free editable map of the world, and one can download and freely use the data. Several studies have looked at the quality of the OSM data in different countries in Europe, see for instance Haklay (2010) for the UK, Girres and Touya (2010) for France and Zielstra and Zipf (2010) for Germany. We are not aware of any such study in the Netherlands, but the dataset is certainly up-to-date and accurate enough since several high-quality datasets of the Netherlands have been recently donated to the OSM project. First, the buildings are coming from the topographic map (Top10NL)‡, and second the roads are very accurate since in 2007 the Dutch mapping company 'Automotive Navigation Data' donated their entire street map of the Netherlands§. Other useful information (e.g.location of bus stops, of airports, of the railways) is also continually added to the OSM database.

The data are organised separately into geometry (polylines) and corresponding attributes (tags). By contrast, most GISs use Simple Features to store geographical objects, so we had to convert the

‡http://wiki.openstreetmap.org/wiki/3dShapes and https://rejo.zenger.nl/inzicht/aanvullende-informatie-over-het-3d-shapes-bestand (in Dutch).
§http://wiki.openstreetmap.org/wiki/AND_Data

datasets to that format before being able to to calculate indicators from it.

5.2 Import and preprocessing of the trajectories

Movement data are usually stored in the GPX format (GPS exchange format), an example of a point stored in the GPX format is shown in the continuation:

```
<trkpt lat="52.196537" lon="5.413356">
  <ele>51.475254</ele>
  <time>2007-03-11T12:50:47Z</time>
  <course>220.490177</course>
  <speed>10.932674</speed>
  <fix>3d</fix>
  <sat>4</sat>
  <hdop>22</hdop>
  <vdop>20</vdop>
  <pdop>29</pdop>
  <quality>1</quality>
</trkpt>
```

The trajectories are imported and stored in the database point by point, and are immediately preprocessed for the different indicators. The computation of some indicators is derived from other indicators (e.g.the maximum speed is calculated after the speeds at all points are derived), hence the preprocessing is done in multiple passes.

5.3 Training of the system

In order for the system to perform well, the proper values for defining the Membership Functions (MF) for every transportation mode have to be defined. Currently this is done in an *iterative process* starting with common senses values for bikes, cars, trains, etc. These values for the MF are encoded in an XML file and the system then classifies the segments. This automatic classification is then compared to the 'ground truth' of the dataset (the results classified by the respondents) and a score is assigned based on the similarity between the automatic and the manual classification. If needed the MF values are adjusted and a new iteration is executed until the results are satisfactory; e.g. the score is above 95%. Note that each iteration is quite expensive as new MFs imply computing again the indicators for the segments and based on the new indicator values the rules are applied for classification. The next excerpt shows a fragment of XML file with MF values:

```
<indicator name="mean_speed">
<mode layer="3" name="walk">
   <values>0,1,8,10</values>
</mode>
```

i.e.the four values of the MF of mean speed for the transportation mode walking (3rd layer) are 0, 1, 8, and 10 km/h.

What is actually going on is a search process to optimal values for the MF functions. For each transport mode there is one MF function described by values. Currently this is a manual trial-and-error process based on analysing the errors.

The end results are extremely sensitive to the determination of the MF parameters. When using a dataset which is not homogenous, many trade-offs should be taken into account due to large differences and possibilities in mode behaviour.

However, this process could also be automated by considering it as a search problem to optimize the score in the high-dimensional space of all values of MF. In principle all steps can be automated: setting four MF values, computing indicators, assigning classification and finally computing the overall score. Hence, we could use one of the well know search/optimization algorithms such as hill climbing, simulated annealing, genetic algorithm, etc. (Russell, 2003; S. Kirkpatrick and Vecchi, 1983; Michalewicz, 1996). As each iteration is relatively expensive, we could try to use a subset of the manual classified segments and optimize for this subset. If good results are obtained we can try these MF values for all manual classified segments. If results are still good then the subset was representative; if the score is not good the subset was not representative. In such a case we should either use a different segment or more segments (e.g.if not all cases were sufficiently represented).

One last warning: the manual classified segments are now considered as the ground truth. For our data this was not really true. We inspected the differences between the manual and the automated classifications and it could be observed that they made about the equal number of mistakes (but different ones); for instance many segments of different public transport modes were classified as one segment, and in general short segments were often not noted. This is something that automatic training (optimizing) of the system can not improve (as it considers the manual classified data as truth and tries to get as close as possible to these classification).

6 Experiments and validation

In order to assess the quality and the applicability of the developed segmentation and classification method, unbiased validation methods have been applied. It was decided to use two different large test datasets for the validations: implying different types of GPS receivers (both new and old), different regions, different timeframes, different sampling periods, and different and non-related organisers of the collection of the involved GPS traces. As explained in Section 5 these two datasets are from the research projects of (1) Bohte and Maat (2009) and (2) Van der Spek *et al.* (2009). The size of the datasets and the variety of all kinds of different movement situations is a guarantee for robust validation.

The datasets have been manually and independently classified within the scope of the original research projects. This classification will now be used as reference material, i.e.'ground truth', for our developed automated approach. There are a few difficulties when comparing the results:

- the segments are not segmented exactly at the same transitions points (so one should accept small differences here),

- our new approach has a more refined classification scheme (hierarchy of transportation modes, see Fig. 2), and
- the manual classification taken as 'ground truth' might contain errors.

In order to cope with the above mentioned validation problems we took the following approach:

- we remapped our refined classification in the validation to the rougher classification of original test data (our classes $x_1, x_2, \ldots$ to class y_1 of the reference classification, e.g.bus and tram are grouped together for the class bus/tram that is defined in the test dataset which was further checked),
- for comparable classes, we collected statistics on the amount of overlap between the two classifications at point level—this resulted in respectively $y_{1_p}, y_{2_p}, y_{3_p}$ agreements and overall y_p of 91.6 %, and
- finally, the individual cases were inspected further where there was a large number of sequential points in a GPS trace that were classified differently. From this inspection (cases) it turned out that there was about an equal amount of manual error and automated errors, i.e.about half of the 8.4% 'error' was indeed correct (4.2%), resulting in an overall score of about 95% (a little lower than 95.8% as there is a small changes that both manual and automated classification are wrong and equal, though we did find no indication for this). Some transportation modes in various situations have been classified with a 100% accuracy, e.g.cars on journeys longer than a few kilometres. When taking into account also the alternative result with the second highest CF, the accuracy was nearly 100%, which is a clear advantage over methods that do not determine values of certainties.

The right side of the Figure 8 shows the GPS data in the spatial extent shown on the left side of the same Figure classified by our fully automated classification and this reinforces the confidence in the above stated positive results of the validation method.

A significant difference between the accuracy of the movement data obtained with relatively old and new devices has not been found. Thanks to the presented approaches, which takes into account the signal shortages, the accuracy is not affected by the deficiencies found in data acquired with older devices. The method can be seen as a worst case scenario where the results can only be improved when newer receivers are used.

The developed method was further checked on various smaller datasets collected by us and others retrieved from the internet for tests with the data acquired with different sampling periods, on a different location and with different GPS devices, as further explained in Biljecki (2010). The segmentation and classification results obtained with the presented automated method are comparable to the above presented findings.

The results may be further improved when optimising the knowledge (membership functions) and using datasets from single sources which contain an uniform behaviour and smaller geographic extent.

7 Discussion and Future Work

The work described in this paper has been initiated by the need of a classification solution for the data acquired for the travel behaviour study conducted by the Department of Urban and Regional Development (Bohte and Maat, 2009) and the Department of Urbanism (Van der Spek *et al.*, 2009) at TU Delft. We have described an algorithm, we have implemented it, and we have validated our results. Their dataset is now classified with a significantly higher accuracy than the existing (semi-manual) method that had been originally used (since we introduced automatic segmentation of the trajectories), and we take into account ten different transportation modes. Apart from travel behaviour research, we believe our method and our prototype is useful in other disciplines since it can be easily extended with new transportation modes (one simply has to define the MFs for all the indicators, i.e.speed, proximity to a certain infrastructure) and the current MFs can also be modified so that they are more realistic for a given country (we have set them up for the Netherlands here). Moreover, the method is universal and we had shown it works for movement data collected with any relatively modern GPS device.

Although it is very difficult to take into account all possible cases in the real-world, the prototype yielded satisfactory results, especially in the segmentation and classification of data containing noise or having a small number of samples.

The errors are usually not caused by the imperfection of the system, rather by specific situations whose modelling would be either complicated or would impair the existing classification performance. We believe that the development of a system that takes into account virtually all possible situations in movement may not be possible.

One reason why we obtained better results than previous attempts is that we rely heavily on the use of geographical data for the indicators, which permits us to resolve gaps, ambiguity between car, bus and tram, and to enrich the trajectories with additional information about the transportation modes. Without geographical data this would not be possible. A few years ago it would have been unthinkable to have (free) access to such datasets, but OpenStreetMap solves that problem and we believe that the quality and coverage of the data uploaded to their servers will only increase in the next year (this is certainly our experience with the data in the Netherlands).

Our method distinguishes between ten transportation modes, which is to the extent of our knowledge, more than any other methods. Although a lot of overlapping characteristics between modes exist, with a careful selection of the indicators and modelling of corresponding membership functions the accurate classification of a large number of modes was made possible. Discerning between car, bus, and tram is done thanks to a developed technique of injecting supplementary confidences based on previous knowledge, which is a novelty. However, there are still cases that we cannot really solve, for instance the difference between bicycles and scooters. In the Netherlands these two modes use primarily the cycling paths, and travel at similar speeds; we tried to use the acceleration to differentiate them but because of the noise in the data that proved unsuccessful. We hope that with newer GPS units that problem will be solved.

For future work, we plan to improve the prototype by training the system with more datasets coming from newer and better GPS units. We plan to allow users to upload their own GPS logs to a

website and let them manually classify their trajectories; that knowledge could then be used to improve our method. Users would also be able to upload their GPS logs and get back segmented and classified trajectories, in a KML file for instance. We also plan on using more auxiliary datasets, such as the type of a road or speed limit. For instance, if it is known that in the Netherlands most of the traffic jams occur on highways, then the classification system may take that fact into account and compensate the speed on such locations (depending on the time of the day). The number of lanes of a road, which is conceptually available in OSM but often not acquired, could be used to alter the MF for the proximity to a road (i.e.a wider road would require a wider MF). Finally, we think that a better classification could be obtained if data about the users were used. Indeed, since a user's trajectory ordinarily contains repetitive journeys, not only in space but also in time and usually with the same transportation mode(s) (e.g.every-day commuting), historical user data may be used to improve the classification in uncertain situations. Similarly, in movement research surveys, extensive data from numerous respondents is often available. By modelling patterns and transportation modes from a group of similar movements, it may be possible to facilitate the classification by searching for a similar trajectory in the database and assign the transportation mode from existing trajectories (classified patterns) in the database.

Acknowledgements

The authors would like to thank the Department of Urban and Regional Development and the Department of Urbanism at the Delft University of Technology for sharing the test datasets.

References

Alkemade, I., Beeldschermkartografie ten behoeve van multi-bron internet GIS. Master's thesis, Delft University of Technology, 2000. .

Asakura, Y., Tanabe, J., and Lee, Y., 2000. Characteristics of positioning data for monitoring travel behaviour. *In*: *7th World Congress on Intelligent Transport Systems, Torino*, Jan., p. 8.

Axhausen, K., *et al.*, 2004. 80 weeks of GPS-traces: Approaches to enriching the trip information. *In*: *Transportation Research Board 83rd meeting*, Jan., p. 28.

Biljecki, F., Automatic segmentation and classification of movement trajectories for transportation modes. MSc Geomatics, GIS technology group, Delft University of Technology, the Netherlands, 2010. .

Bohte, W. and Maat, K., 2008. Deriving and Validating Trip Destinations and Modes for Multi-day GPS-based Travel Surveys: An Application in the Netherlands. *In*: *Transportation Research Board 87th Annual Meeting*, p. 17.

Bohte, W. and 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. *Transport Res C-Emer*, 17 (3), 285–297.

Bohte, W., Maat, K., and Quak, W., 2008. A method for deriving trip destinations and modes for GPS-based travel surveys. *In*: J. Van Schaick and S. Van der Spek, eds. *Urbanism on Track*. IOS Press, chap. 10, 129–145.

Bricka, S. and Bhat, C., 2006. Comparative Analysis of Global Positioning System-Based and Travel Survey-Based Data. *Transportation Research Record: Journal of the Transportation Research Board*, 1972, 9–20.

Buchanan, B.G. and Duda, R.O., 1982. Principles of Rule-Based Expert Systems. *In*: M. Yovitz, ed. *Advances in Computers*., Vol. 22 Academic Press, New York, p. 62.

Byon, Y.J., Abdulhai, B., and Shalaby, A., 2009. Real-Time Transportation Mode Detection via Tracking Global Positioning System Mobile Devices. *Journal of Intelligent Transportation Systems*, 13 (4), 161–170.

De Boer, A., Analysis of GPS logs for algorithm design of movement behavior studies. , 2008. , Technical report, Delft University of Technology.

Dodge, S., Weibel, R., and Forootan, E., 2009. Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. *Computers, Environment and Urban Systems*, 33 (6), 419–434.

Draijer, G., Kalfs, N., and Perdok, J., 2000. Global Positioning System as data collection method for travel research. *Transportation Research Record: Journal of the Transportation Research Board*, 1719, 147–153.

Girres, J.F. and Touya, G., 2010. Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, 14 (4), 435–459.

Gonzalez, P., *et al.*, 2010. Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks. *In*: *IET Intelligent Transport Systems*, Vol. 4, Jan., 37–49.

Grazia, C.U., Introduzione ai sistemi esperti. , 2006. , Technical report, Sistemi di elaborazione delle informazioni, Dipartimento di Scienze, Università degli Studi "Gabriele D'Annunzio".

Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 37 (4), 682–703.

Holzmann, C.A., *et al.*, 1999. Expert-system classification of sleep/waking states in infants. *Medical and Biological Engineering and Computing*, 37 (4), 466–476.

Kotte, I., Een kartografisch expert systeem ten behoeve van presentatie van gedistribueerde geografische informatie. Master's thesis, Delft University of Technology, 2002. .

Lester, J., *et al.*, 2008. MobileSense - Sensing Modes of Transportation in Studies of the Built Environment. *In*: *International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems - UrbanSense08*, Oct., p. 5.

Liao, L., *et al.*, 2006. Building personal maps from GPS data. *Annals of the New York Academy of Sciences.*, Vol. 1093, 249–265.

Liao, L., *et al.*, 2007. Learning and inferring transportation routines. *Artificial Intelligence*, 171, 311–331.

Maat, K. and Timmermans, H., 2006. Influence of Land Use on Tour Complexity: A Dutch Case. *Transportation Research Record: Journal of the Transportation Research Board*, 1977, 234–241.

McGowen, P. and McNally, M., 2007. Evaluating the Potential To Predict Activity Types from GPS and GIS Data. *In*: *Transportation Research Board 86th meeting*, Jan., p. 21.

Michalewicz, Z., 1996. *Genetic Algorithms + Data Structures = Evolution Programs*. 3rd edition, Springer-Verlag, Berlin.

Ministerie van Verkeer en Waterstaat, Mobiliteitsonderzoek Nederland. Het onderzoek. Technical report, 2009. .

Mountain, D. and Raper, J., 2001. Modelling human spatio-temporal behaviour: a challenge for location-based services. *In*: *GeoComputation - Brisbane*, p. 9.

Nickles, M. and Sottara, D., 2009. Approaches to Uncertain or Imprecise Rules - A Survey. *In*: G. Governatori, J. Hall and A. Paschke, eds. *Rule Interchange and Applications.*, Vol. 5858 of *Lecture Notes in Computer Science* Springer Berlin / Heidelberg, 323–336.

Ranjitkar, P., *et al.*, 2002. Car-Following Experiments Using RTK GPS and Stability Characteristics of Followers in Platoon. *In*: *Proceedings of 7th International Conference on Application of Advanced Technologies in Transportation Engineering*, Vol. 245, Jan., 608–615.

Rearden, P., *et al.*, 2007. Fuzzy Rule-Building Expert System Classification of Fuel Using Solid-Phase Microextraction Two-Way Gas Chromatography Differential Mobility Spectrometric Data. *Analytical Chemistry*, 79 (4), 1485–1491.

Reddy, S., *et al.*, 2008. Determining transportation mode on mobile phones. *In*: *12th IEEE International Symposium on Wearable Computers*, 25–28.

Reddy, S., *et al.*, 2010. Using Mobile Phones to Determine Transportation Modes. *ACM Transactions on Sensor Networks*, 6 (2), 13–40.

Russell, S., 2003. *Artificial Intelligence*. Englewood Cliffs: Prentice Hall.

S. Kirkpatrick, C.D.G. and Vecchi, M.P., 1983. Optimization by Simulated Annealing. *Science*, 220 (4598), 671–680.

Schüssler, N. and Axhausen, K.W., 2009. Processing Raw Data from Global Positioning Systems Without Additional Information. *Transportation Research Record: Journal of the Transportation Research Board*, 2105 (4), 28–36.

Shalaby, A., *et al.*, 2006. New Tools for GPS-based Travel Surveys and Traffic Monitoring. *In*: *New Frontiers in Transport Systems*, p. 33.

Shortliffe, E. and Buchanan, B.G., 1975. A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23 (3-4), 351–379.

Shortliffe, E., *et al.*, 1975. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. *Computers and biomedical research*, 8, 303–320.

Spaccapietra, S., *et al.*, 2008. A conceptual view on trajectories. *Data & Knowledge Engineering*, 65, 126–146.

Stopher, P., *et al.*, Deducing mode and purpose from GPS data. , 2008. , Working paper ITLS-WP-08-06, Institute of transport and logistic studies, The Australian Key Centre in Transport and Logistics Management, The University of Sydney.

Stopher, P., FitzGerald, C., and Zhang, J., 2007. Search for a global positioning system device to measure person travel. *Transport Res C-Emer*, 16, 350–369.

Van der Spek, S., *et al.*, 2009. Sensing Human Activity: GPS Tracking. *Sensors*, 9, 3033–3055.

Van der Spek, S., 2010. Tracking Tourists in Historic City Centres. *In*: U. Gretzel, R. Law and M. Fuchs, eds. *Information and Communication Technologies in Tourism 2010. Proceedings of the International Conference in Lugano, Switzerland, February 10–12, 2010* Springer Vienna, chap. 5, 185–196.

Van Oosterom, P., 1999. Rule-based Polygon Classification of Topologically structured Topographic Data converted from Spaghetti Data. *In*: *Computational Cartography, Dagstuhl-seminar, 19-24 october 1999*, Oct.

Van Oosterom, P., *et al.*, 2001. Multi-Source Cartography in Internet GIS. *In*: *Proceedings 4th AGILE Conference, Brno*, April., 562–573.

Verbree, E., *et al.*, 2005. GPS-monitored itinerary tracking: Where have you been and how did you get there?. *Geowissenschaftliche Mitteilungen*, 74, 73–80.

Wentz, E.A., *et al.*, 2008. Expert system classification of urban land use/cover for Delhi, India. *International Journal of Remote Sensing*, 29 (15), 4405–4427.

Wolf, J., 2000. Using GPS data loggers to replace travel diaries in the collection of travel data. Thesis (PhD). Georgia Institute of Technology.

Wolf, J., Guensler, R., and 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.

Zhang, J., *et al.*, 2006. On the relativistic Doppler effect for precise velocity determination using GPS. *Journal of Geodesy*, 80, 104–110.

Zheng, Y., *et al.*, 2008a. Understanding mobility based on GPS data. *In*: *Proceedings of ACM conference on Ubiquitous Computing (UbiComp 2008), Seoul, Korea*, 312–321.

Zheng, Y., *et al.*, 2010. Understanding Transportation Modes Based on GPS Data for Web Applications. *ACM Transaction on the Web*, 4 (1), 1–36.

Zheng, Y., *et al.*, 2008b. Learning transportation mode from raw GPS data for geographic applications on the web. *In*: *International World Wide Web Conference: Proceeding of the 17th international conference on World Wide Web*, Apr., 247–256.

Zielstra, D. and Zipf, A., 2010. A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. *In*: *Proceedings 13th AGILE International Conference on Geographic Information Science*, Guimarães, Portugal.