On the requirements on spatial accuracy and sampling rate for transport mode detection in view of a shift to passive signalling data

On the requirements on spatial accuracy and sampling rate for transport mode detection in view of a shift to passive signalling data

O. Burkhard[a,*], H. Becker[b], R. Weibel[a], K. W. Axhausen[b]

[a] *University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland*
[b] *ETH Zurich, Stefano-Franscini-Platz 5, 8093 Zurich, Switzerland*

Abstract

GPS based campaigns have been hailed as an alternative to transportation surveys that promise relatively high accuracy at a relatively low burden on the participants and fewer forgotten trips. However they still necessitate the recruitment of participants and are thus potentially biased and certainly not encompassing significant parts of the population. Given the high penetration of mobile phones, passive tracking by telephone providers would alleviate those two shortcomings at the cost of reduced sampling frequency and positional accuracy. The trade-off in quality has not yet been quantified and therefore recommendations on sensible thresholds are not yet available. In this study therefore, instead of presenting yet another method for mode of transport classification, we therefore compare the performance of existing mode detection schemes under deteriorating sampling rates and positional accuracies. As a possibility to compensate for the deteriorating signal we also calculate features from users' positional histories that could be beneficial if their behaviour is repetitive. The evaluation is not only based on pointwise accuracy, but includes quality measures that pertain to trips as a whole. We find that the necessary accuracy and sampling rate for applications will depend on whether the information of whole trajectories can be used, or whether only the current information is available. The former being relevant to ex-post analyses while the latter situation appears more frequently in near-time analyses. For segmentwise classification, there is no major impact on the quality of the classification by the tested levels of spatial accuracies as long as the sampling intervals can be kept at or below a minute, whereas for point based classification the sampling interval should be between 30 seconds and a minute and increasing spatial accuracy always improves the classification.

Keywords

Transportation Mode Detection, Data Quality, Passive Tracking

1. Introduction

Understanding different aspects of the travel behaviour of a population has long been the focus of study for scholars in the transportation sciences. Increasingly however, this research is drawing interest from academics in other fields that use those transportation science tools to solve their problems. Those can be very diverse and range from predictive policing (Leuzzi et al., 2017) to modelling vehicular emissions in a city (Nyhan et al., 2016) and monitoring health (Saeb et al., 2016).

There are several ways to obtain the desired information. Traditionally, questionnaires were sent to a relatively small random sample of the population, in which the participants were asked about the above mentioned information regarding their travel behaviour concerning, for a long time, one day only (Axhausen et al., 2002). Surveys have the advantage of being semantically rich: The questionnaire can contain very

*Corresponding author
Email address: `oliver.burkhard@gmail.com` (O. Burkhard)

detailed questions about the motivations that are so far only available by asking people directly. On the other hand there are also shortcomings to this method, which typically include a low response rate, an uneven response rate between the different strata of the population, forgotten trips, the relatively high burden on the participants, and the price of obtaining the data that grows linearly with the sample size (Bricka et al., 2009; Bricka & Bhat, 2006; Furletti et al., 2013).

About a decade ago new types of positioning data, such as GPS became broadly available, and more recently new mobile phone standards that allow even better positioning (Leuzzi et al., 2017; Dammann et al., 2015). The ever increasing availability of such technologies has prompted research on methods to answer traditional questions of the transportation sciences such as trip chains (Jiang et al., 2017), mode and purpose of trips (Zolliker et al., 2015), and OD-matrices (Ni et al., 2018).

These new sources of data have been used to overcome the aforementioned shortcomings. For this, either a dedicated logger of a global navigation satellite system (GNSS), of which GPS is the most prominent representative, or a mobile phone on which a GNSS-logging app is run (Shen & Stopher, 2014) is given to the participants of the study. They produce (at least) a set of position fixes in the form of $(x_{u,t}, y_{u,d}, z_{u,t}, u, t)_{u,t}$ for different users u and time stamps t. GNSS based surveys promise fewer forgotten trips (Nitsche et al., 2012) while maintaining an impressively high – albeit not uncontested – accuracy (Shen & Stopher, 2014; Stenneth et al., 2011; Prelipcean et al., 2017). However, they are semantically poor and the recruitment cost still grows linearly in the number of participants.

One way of overcoming the limitation of linearly growing cost of obtaining the information is using passive tracking of mobile phones. Telephone companies routinely collect different kinds of information about their users. At a minimum they collect call detail records (CDR) which contain the mobile phone tower that was used to transmit an incoming or outgoing text message, phone call or use of mobile data, which can be used to infer information about the mobility of the users, as for example shown by Bachir et al. (2019). This level of data, however, suffers from quite significant shortcomings in terms of spatial accuracy and temporal granularity (Burkhard et al., 2017). A much more useful kind of data that can be collected passively from mobile phones, however, are the trilaterated position updates that can be sampled at frequencies that are much higher than what can be obtained from CDR and obviously causally linked to movement of the phone. While collecting this type of data comes at an additional cost to mobile phone network operators, they are increasingly inclined to do so (Musolesi, 2014; Leber, 2013). However, this data still comes with spatial uncertainty (Müller et al., 2016) and is unlikely to be collected at temporal granularities usually seen for GNSS based studies. There is thus a gap between the kinds of data typically used for passive tracking and the kinds typically used for transportation mode detection. Exploring where within this gap the methods used on actively tracked data seize to be useful is on open question that this study addresses. For the temporal parameter alone, sparser sampling rates have already been successfully applied by e.g. Bolbol et al. (2012), but interaction effects may be present and have to be tested for.

Passive tracking data have one disadvantage though, compared to positioning data obtained from a GNSS, especially if the latter is collected on a mobile phone: Passive tracking only obtains position updates and misses out on the instantaneous speed that is available on all GNSS trackers, and also on accelerometry, the magnetic field, and other information available through mobile phones. Nevertheless, passive positioning data holds the promise of removing two significant limitations shared most often by studies based on recruited users: First, the data would be of longitudinal nature, allowing for a whole new range of questions to be answered. Second, the scope of studies, both spatially as well as in the number of users under study, could be significantly larger, as the data is collected anyway in he background, and there is no need to recruit users specifically. Data privacy is of course a major concern in such studies and has to be handled with extreme care, as both very private information on large parts of the population as well as the reputation of the data providing telephone company are at risk.

As promising as passively tracked mobile phone location updates sound, it remains unclear how close to GNSS data they need to be in order for the GNSS-based algorithms for transport mode detection to be transferable to that new, passive type of data, especially since most of the evaluation so far has been based on summary statistics over the population at large, since ground truth labels were not available (Huang et al., 2019). In particular it is relevant to know at which *spatial accuracy* and *temporal granularity* the

algorithms break down. Having those thresholds would provide the limits that mobile phone companies have to achieve in order for their data to be useable for transport mode detection. Establishing these thresholds is precisely the purpose of this study and can be seen as moving towards addressing the Problems 4 and 6 outlined in Huang et al. (2019) in that on the one hand, having six different modes of transportation are on the upper end of what can be found for passive tracking, and on the other hand the evaluation happens not on the basis of summary statistics but on actual labels on the individual trips. This then allows not only the coarse OD-Matrix style of analysis often seen in studies relying on call detail records (Csáji et al., 2013) but to engender more detailed analyses in the vein of Rinzivillo et al. (2014).

For our study we use a longitudinal sample of 138 users over a duration of 6 months that contains GNSS points as well as ground truth labels on mode of transport. The data was collected on the normal daily routines of the participants and thus the data collected can be thought of as being close to what passive tracking would yield in terms of behaviour and can thus be seen as realistic. Given the users were not researchers but members of the public that bought into a mobility as a service scheme described in Section 3.1, their ground truth labels may not be as accurate as those gathered by dedicated researchers with a personal interest in labelling quality, but on the other hand the daily routines are not restricted to those of researchers, providing a somewhat broader and more realistic scope.

On this data set we apply methods that were successfully used on GNSS data to infer the mode of transport, while gradually deteriorating both the spatial accuracy and temporal granularity, and thus getting closer to the kind of data one has to expect from passive tracking by mobile phone companies. This allows us to see how the quality of the prediction declines with the decreasing quality of the signal. To correctly account for the differences of passively tracked mobile phone locations with respect to GNSS – no information on speed and acceleration, but longitudinal data are available – we also allow features taken from the longitudinal nature of the data to be used in the classification.

The contributions of this work are thus:

- An analysis of how popular GNSS-based transportation mode detection algorithms perform under deteriorating conditions of both spatial accuracy and temporal granularity of the underlying data.
- Recommendations derived from that analysis on the necessary data quality for purely passive tracking of mobile phone users.
- An assessment of different transportation mode detection techniques on data that has breadth in the user base but less than perfect labelling.

2. Related Work

2.1. Problem definitions

The problem of transport mode detection is not clear-cut: Depending on the application of the resulting classifiers they are being used to answer different questions (Prelipcean et al., 2016). For a detailled definition of the terminology used here, please refer to Section 3.5. There are two major approaches to solving the problem: pointwise and segmentwise classification.

In pointwise classification, as shown on the left hand side of Figure 1, the input consists of a sequence of feature vectors – both directly sensed and derived – that are related to the individual measurements by a GNSS sensor that are individually classified (Bolbol et al., 2012). The misclassification of a single point in an otherwise correctly identified stage will not change the overall number of correctly classified points (and thus accuracy, recall and the F1 score) by much, but will lead to the introduction of an additional stage, which can be problematic, if the user is interested in the composition of entire trips (Prelipcean et al., 2017).

There are several alternative strategies to remedy this problem. The simplest one, often used only implicitly, is to calculate features that are based on other measurements in a time window around the point in question (Stenneth et al., 2011). This is illustrated by the left two braces in Figure 1. As temporally close points have similar surroundings, also the derived features should be similar and hence there is a higher probability of the same label being assigned. Alternatively, also a fixed number of neighbouring fixes

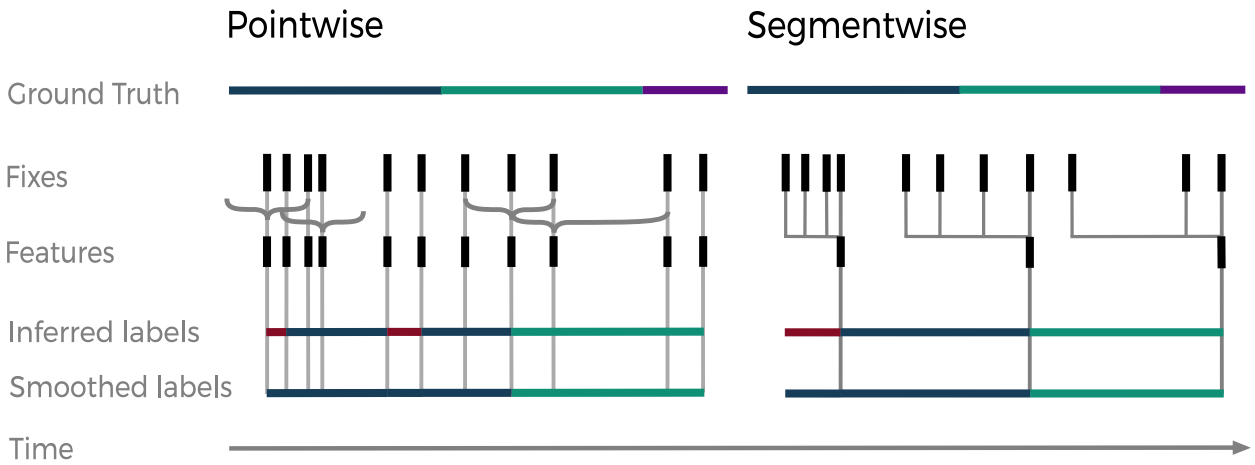


Figure 1: Two major approaches of transport mode classification: segmentwise and pointwise. A detailed description can be found in the text.

could be taken as the basis of feature calculations, as shown by the right two braces in Figure 1, where the neighbourhoods have the same number of fixes, but span different durations.

Another way to combat errors in pointwise classification is to smooth the labels after classification. Essentially *small* numbers of differing labels within a neighbourhood of another label get changed to that majority label, as shown in the bottom part of Figure 1. Different smoothing schemes are possible but most studies seem to be using ones based on either hidden Markov models (Reddy et al., 2010; Nitsche et al., 2012, 2014; Shah et al., 2014) or a simple majority vote in a moving window (Prelipcean et al., 2016).

The alternative to pointwise classification is segmentwise classification, shown on the right hand side of Figure 1. There, a trip is first split into segments. The features are then calculated for those segments as a whole and used in the classifiers. Most commonly, this pre-segmentation is performed using episodes of (near) zero velocities or rules identifying gaps in the data (Chen et al., 2010; Gong et al., 2012; Huss et al., 2014; Pereira et al., 2013; Sauerländer-Biebl et al., 2017; Zhang et al., 2012). Multiple segments with the same inferred travel mode are then combined to a stage.

Alternatively, in some studies, the trips or stages to be classified are delineated by the users themselves, which simplifies classification considerably (Bohte & Maat, 2009; Bolbol et al., 2012), but limits the use of such methods to cases where the people of whom the data are being analysed can give that kind of feedback.

One method that falls somewhere between pointwise and segmentwise classification are conditional random fields (CRFs) on pointwise features. While the classification is clearly pointwise, the fact that this method typically learns that the same label repeats itself leads to longer sequences of identical labels. CRF's have been found to perform worse than two-stage approaches (Zheng et al., 2008). Other alternatives, such as recurrent neural networks (Lin et al., 2017), sequence to sequence models (Sutskever et al., 2014) or attention based classifiers (Vaswani et al., 2017) are at least conceptually very well suited for the task, but have so far not been the focus of research.

Another factor that can drastically simplify the problem of mode detection, as pointed out by for instance Huang et al. (2019) is the question which and how many modes to include in the study, with many studies in passive tracking contenting themselves with separating modes that are relatively easy to detect.

2.2. Features

Once the exact object that needs to be classified is determined, the next question is of course what features are to be used.

As most studies are GNSS-based and such data come with an estimation of velocity and its derivatives such as averages (Bohte & Maat, 2009; Gonzalez et al., 2008; Schuessler & Axhausen, 2009; Stenneth et al., 2011; Stopher et al., 2008; Zhang et al., 2012; Zheng et al., 2008), extremes (Bohte & Maat, 2009; Gonzalez

et al., 2008; Nitsche et al., 2012, 2014), quantiles (Gong et al., 2012; Huss et al., 2014; Nitsche et al., 2012, 2014), acceleration (Huss et al., 2014; Schuessler & Axhausen, 2009; Shafique & Hato, 2015; Zhang et al., 2012; Zheng et al., 2008) and variability in speed (Zheng et al., 2008) are by far the most common features for transport mode detection. This is of course very sensible, as speed is also very capable of distinguishing certain modes such as walking vs. taking the train.

Still remaining with GPS, there are several studies that incorporate other information that can be gathered directly from the GNSS sensor, such as the precision of the signal (Ellis et al., 2014; Gong et al., 2012; Stenneth et al., 2011; Xiao et al., 2015) or the number of satellites in view (Gong et al., 2012; Jahangiri & Rakha, 2015).

Besides GPS-information there is a range of studies using other spatial information, such as static GIS information which most often takes the form of proximity to public transport stations or lines (Chen et al., 2010; Gong et al., 2012; Moiseeva & Timmermans, 2010; Semanjski et al., 2017; Stenneth et al., 2011; Zhu et al., 2016), but can also be the kind of road used (Semanjski et al., 2017). Other studies have managed to incorporate dynamic information on public transport, such as the real-time location of bus coaches and trains (Stenneth et al., 2011).

Departing from spatial information, some studies rely entirely on non-spatial sensors such as accelerometers (Eftekhari & Ghatee, 2016), which can be used to preserve the privacy of the participants. This is highly desirable in situations where positional information is not at the centre of interest. However, since the use case in the setup of this paper is passive tracking in the context of traffic surveys, where positional information *is* of interest and those sensors are not available in that situation, we will not discuss them further in this article.

Some studies have incorporated information or preferences of the users (Moiseeva & Timmermans, 2010; Stopher et al., 2008), but this has the disadvantage of generalising poorly to situations where this information is not available.

2.3. Classifiers

Many classifiers have been used for the task of classification in mode detection. The ones that are most closely tied to the specificities of the problem at hand are the rule-based classifiers that typically have relatively rigorous boundaries on a relatively small number of features (Bohte & Maat, 2009; Chen et al., 2010; Gong et al., 2012; Sauerländer-Biebl et al., 2017; Schuessler & Axhausen, 2009; Stopher et al., 2008; Marra et al., 2019). In situations where there are more features affecting the classification, support vector machines (Bolbol et al., 2012; Pereira et al., 2013), decision trees (Reddy et al., 2010), and random forests (Ellis et al., 2014; Mäenpää et al., 2017) are the most popular choices. They have the advantage of performing quite well while being relatively easy to implement. Among those papers that tried to classify in a pointwise fashion while still retaining reasonable overall trips, hidden Markov models are clearly a popular approach (Bantis & Haworth, 2017; Nitsche et al., 2014; Reddy et al., 2008).

A substantial part of earlier research has been content with presenting one strategy that worked in the context of their research questions. This is typically the case where detection of transport mode is not the primary goal but only a necessary step on the way (Gong et al., 2012; Huss et al., 2014; Schuessler & Axhausen, 2009).

For passive tracking, unsupervised methods have also found their application, as presented by Huang et al. (2019). While unsupervised approaches can generate important insights, no precise quality measures are possible (Chin et al., 2019), which is why they will not be further considered in this article.

2.4. Evaluation metrics and collected data

In most cases, just precision, recall, accuracy and/or the F1 score are being reported (often by transport mode). However, especially in the case of unbalanced data (i.e. vastly different frequencies for the different modes) a high accuracy does not necessarily mean good classification. Therefore some authors have also provided either Cohen's Kappa (Bolbol et al., 2012; Huss et al., 2014) or the Chi-Squared (Bantis & Haworth, 2017). Other metrics, such as the ones proposed by Prelipcean et al. (2016) have yet to be widely adopted.

Also, different sampling frequencies adversely affect the comparability of the studies. While GPS-based features are relatively often sampled at 1Hz there are quite a number of researchers who sample at lower

rates down to 1 data point a minute (Bantis & Haworth, 2017; Bolbol et al., 2012; Mäenpää et al., 2017). While schemes for battery conservation have been devised (Linnap & Rice, 2014) they typically seem to be contemplated before the study, i.e. without knowing what would actually be sufficient data for the task. In the spatial domain, however, the accuracy of the GPS was usually treated as a given and deteriorations of this signal have not widely been considered.

2.5. Comparison to results from the literature

It is not a trivial task to compare the different methods proposed in the literature. Many of them use information that is not available in the situation of passive tracking by telephone companies, such as instantaneous speed, measurements from accelerometers or information on vehicle ownership (Stenneth et al., 2011; Nitsche et al., 2014; Bantis & Haworth, 2017; Feng & Timmermans, 2016).

A further problem that often is insufficiently addressed is a clean split into training and testing data. Such a split may not happen at all (Bohte & Maat, 2009; Bolbol et al., 2012; Gong et al., 2012; Chen et al., 2010), which may yield overly optimistic results as the methods are optimised to the data used for evaluation and the out-of-bag error may be larger. Even if the training and testing data are split, there may be an overlap between the two either in terms of moving windows on which features are calculated (Feng & Timmermans, 2016) or in terms of the users (Bantis & Haworth, 2017; Semanjski et al., 2017; Stenneth et al., 2011; Zhang et al., 2012; Xiao et al., 2015). This is not to say that a split according to users, such as done by Huss et al. (2014) is the only way. However, the generalisation from training to validation should be the same that can be expected from training to the actual application. Thus, if the trained classifier will only be applied on the users for which some labels are available, using different users for validation as for training is not necessary. A selection demonstrating the diversity of different problems that are being solved under the name of mode detection can be found in Table 1. What this means is that it is impossible to tell whether the results obtained by the method leading to the 90% reported by Bantis & Haworth (2017) are better or worse than the ones leading to the 94% reported by Semanjski et al. (2017), even if in this instance the authors of these two papers at least both report on the same measure (overall accuracy). This in turn implies that while overviews comparing results as is done in, for instance, the latter paper can be of indicative value, the comparison cannot be completely fair. As an example of this problem we illustrate just how strong a seemingly small change in the problem statement can affect the results in Section 4.4. To allow a fair comparison given the fracturing of research in terms of the problem definition, the people under study and the available information, we need to compare the important approaches on the same dataset and using the exact same problem statement. Of course, it would be best if a generally accepted dataset and problem definition were available on which the community would agree to evaluate all their ideas, but in want of that in this paper a comparison of common approaches must suffice. In this paper we therefore show the behaviour of a multitude of approaches in the presence of deteriorating data quality.

3. Materials and Methods

Our goal is to demonstrate how close to a GNSS signal passively sensed data needs to be in order to allow transport mode detection using the techniques typically associated with GNSS data. For this, we progressively distort the GNSS data spatially and subsample it temporally in order to find the limits at which the traditional GNSS approaches to transport mode detection are no longer useful.

3.1. Data

The data was collected as part of a pilot study for a new *Mobility as a Service* (MaaS) offering of the Swiss federal railway company SBB (Becker et al., 2018) with 138 participants across Switzerland. They were selected to cover a variety of living conditions so that the experiment covered the whole country, and different types of spatial backgrounds (from small village dwellers to residents of large cities). The MaaS offering included both unlimited use of public transportation within Switzerland as well as the lease of an electric car (in addition to the cars and bicycles already available in their household). Hence, respondents showed a diverse and highly multi-modal travel behaviour.

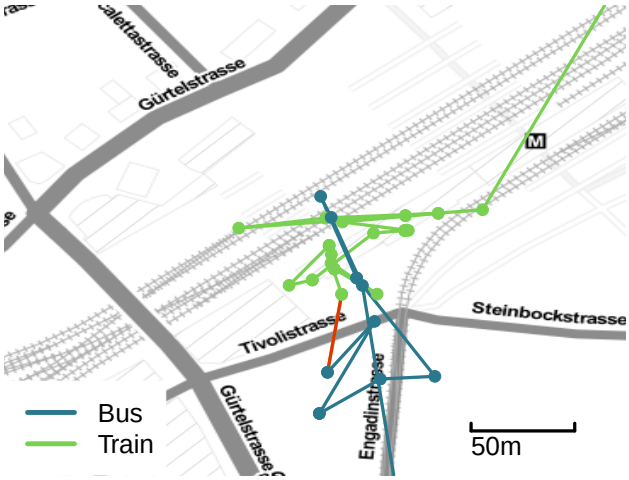


Figure 2: Detail of an annotated trajectory near the main station in the city of Chur. The red line highlights two consecutive points with different not-walk modes of transportation without a walk stage between them. Here the situation of the GNSS fixes would suggest a rather stationary behaviour indicative of a walk stage.

As part of the study, each participant had to record a travel diary for the whole duration of the study (about one year) using a re-branded version of the MotionTag smartphone-app.[1] The app uses the device's location services to record coordinates, transmits them to a server, where the records are classified into trips and activities using proprietary algorithms. Respondents were able to review their records from within the app and were asked to confirm or edit activity types and modes. Users were not able to alter the geometry of trips or the segmentation by themselves, but were encouraged to report any erroneous records. The sampling frequency was set to 1Hz, but was of course heterogeneous due to e.g. signal loss in trains or tunnels.

A partial dataset containing all records and trips made between March and August 2017 (6 months) was available for this research. The data contain both the raw data (*waypoints*) and the annotated trip data (*tracks*):

- *waypoints*: 21,119,962 observations (after subsampling to the highest temporal granularity used here, removing the most obviously wrongly labelled trips and using only trips that remained within Switzerland) containing user ID, timestamp, longitude, latitude;
- *tracks*: 117,091 observations containing user ID, start date/timestamp, end date/timestamp, distance, PostGIS geometry, detected mode[2], confirmed mode, user comment. 96.1 % of the tracks have a user-confirmed transport mode.

The user-confirmed trips from the *tracks* dataset represents the ground truth, the *waypoints* dataset is considered the raw data.

The tracking system was not always able to correctly identify access and egress walk stages. This gets reflected in the most common sequence labels in the dataset shown in Table 2: For example the sequence "Train" appears about twice as often as the (more likely) sequence "Walk, Train, Walk".

One example of this can be seen in Figure 2, where the last train point is immediately followed by a bus fix, even though there clearly seems to be a time where the user in question was moving very slowly in the train station area. This limitation has to be accounted for when interpreting the results.

In Table 2 the unweighted counts are listed. Since some modes usually have longer travel times than others, we also provide the number of points (on the 5 second granularity) in Table 3. A second limitation of the ground truth data can be seen in Figure 3.

[1] `https://motion-tag.com`

[2] The options were: airplane, boat, coach, bus, tram, train, car, bike, ski and walk. We removed trips containing the extremely rare modes of ski, coach, airplane and boat.

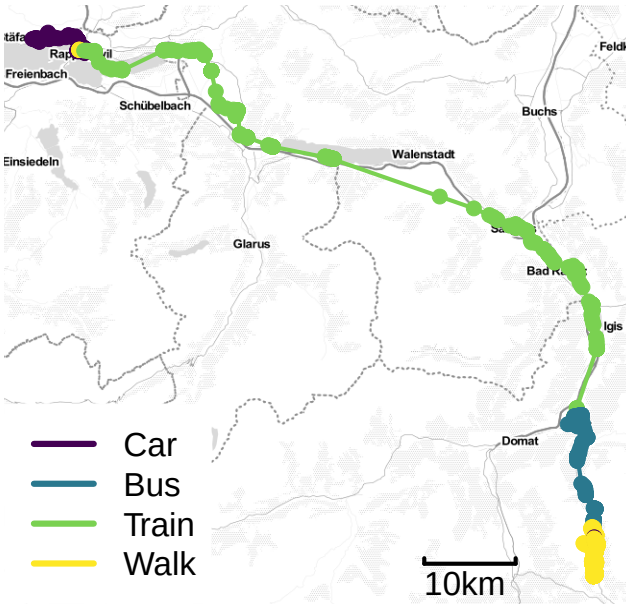


Figure 3: Example trajectory of one user for one day. The train stage of the trajectory is clearly most affected by GNSS signal loss. Other modes of transportation seem to suffer less, particularly walking usually has good coverage.

Since GNSS signal loss is most common on trains, the actual number of hours travelled in trains was higher than suggested by the point counts. It is evident that the data is heavily skewed in favour of the *Train* and *Car* modes, which has to be accounted for when training the classifiers lest one obtains too optimistic results.

The geographical distribution of the data points is presented in Figure 4. The municipalities with most recorded data lie along the main traffic routes of the country, whereas the more rural areas feature fewer hours of recordings. The collection covers the entire country with the exception of the Alps.

Areas that are not considered habitable by the Swiss Federal Office of Statistics [3] (lakes, glaciers and rock) are not considered and left blue (lakes) or in hillshading (glaciers and rock).

3.2. Subsampling of the data

For the subsampling, a relatively pragmatic approach was used. Time was partitioned into episodes of equal length (e.g. 5 minutes) and for each of those episodes, the first observation was taken. This ensures that the retained measurement is a true measurement and not the result of an aggregation (e.g. mean or median), which would not be a fair comparison. Given that the original sampling rate was about 1 Hz, the unevenness of the time differences between the retained points should not be materially increased by the simplicity of the scheme.

This regime does not affect the parts of the data where the original signal was already coarser, due to e.g. signal losses in trains and tunnels. The highest temporal resolution was chosen to be 5 seconds which is considerably higher than what seems to be available for passive data today, but may become a reality with ultra dense 5G networks (Koivisto et al., 2017). The lowest temporal resolution was chosen to be 5 minutes, which is in the order of magnitude of individual short trips. Including even lower levels of temporal resolution would lead to problems when identifying the trips as such. Yet, data quality does not appear to be a problem. In recent statements, Swisscom, the largest Swiss mobile phone network provider, reported to collect 20 billion events per day (Rollier, 2015). With 6.6 million customers as of 2018 (Swisscom AG, 2018), this corresponds to a 30 second resolution on average. Given a certain number of actually unused

[3] `https://www.bfs.admin.ch/bfs/en/home.html`

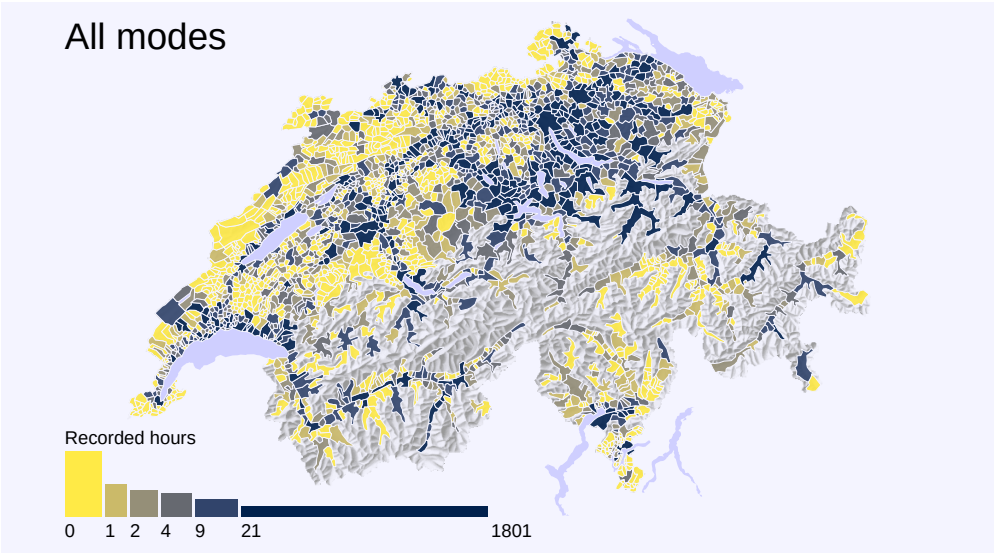


Figure 4: Distribution of the data points within Switzerland. The inhabitable areas of municipalities are colour-coded by how many hours worth of data was collected in them, with the colour code representing the quantiles of municipalities. There were much more recorded hours in larger cities, especially Zurich, which is to be expected. Corresponding maps by mode of transport can be found in the Supporting Materials.

phones, the actual sampling rate may be even higher. Table 4 illustrates the subsampling used (assuming 5 minute intervals).

3.3. Distorting the data

As spatial distortion we added jointly normally distributed, uncorrelated pseudo-random errors to the position obtained from the GNSS signal. The position to be distorted was not latitude and longitude directly, but the x and y coordinates in the EPSG 2056 reference system that correspond to the positions obtained through GNSS. We used 0 m, 25 m, 50 m and 100 m as standard deviations. The upper bound reflects the order of magnitude to be expected from current LTE trilateration (Müller et al., 2016). From there we go all the way down to 0, i.e. the uncertainty we currently have with GNSS's. The upcoming positioning data from 5G promise to be even more accurate than GNSS's (Koivisto et al., 2017), but the effects of that cannot yet be tested in the setup of this study.

3.4. Methods

The focus of our work lies on the effects of the deterioration of the positioning signal and not primarily on the merits of one transportation mode detection method over another. Therefore, lest our results be consequence of idiosyncrasies of a certain combination of overall classification strategy (pointwise, smoothed pointwise or segmentwise) and classifier, respectively, we try to cover most relevant cases. As for the features, we restrict ourselves to those that nowadays can be assumed to be available for passively sensed data.

3.5. Terminology

As already stated, we start with individual *fixes* that denote information (such as the position, but also includes features such as proximity to public transport) at a given point in time.

A set of temporally contiguous fixes that are semantically very close can be grouped into a *segment*. However, there are competing notions of a segment, reflecting different ontologies of movement. A set of contiguous fixes that shares a common mode of transport is called a *stage* or an *inferred stage*, depending on whether we use the ground truth or the inferred labels as the basis for the grouping.

On the other hand we call *segment* a set of contiguous fixes that share characteristics pertaining to the displacement over time implied by the fixes. Those segments are used in some of our classification approaches to yield more stable or more accurate results. The term "segmentwise classification" always refers to a classification based on that type of segments, as at the time of classification the delineation of the stages is unknown.

Both kinds of division of trips are partitions of the fixes used, i.e. every fix belongs to exactly one segment and exactly one stage. However, the segments and the stages need not coincide. For example, it seems very possible that a bus stage comprises many move segments (one for every move between two consecutive bus stops) whereas it will only be a single stage, as the mode of transport is *Bus* for the whole journey.

Finally a *trip* denotes the smallest contiguous set of fixes that are deemed a journey between two places where the person performed some meaningful activity (e.g. a trip between home and work). The splitting of the raw data into trips is not part of this work, as the stays were for the most part very clearly discernible. This assumes long stays with short times of movement in between, as was already observed in other studies (Burkhard et al., 2017).

3.6. Features

For the "raw signal", we use the distorted and subsampled data mentioned in Section 3.3. To this raw signal, we added information that would be available to any service using passively sensed positioning data in many countries: Speeds that are calculated on consecutive position fixes, distances to public transport facilities relevant to the study area and quantiles thereof over moving windows. The relevant public transport facilities were both the stops (point data) and routes (line data) of buses, trams and trains.

The positions of the public transport stops were obtained through the open data portal for public transport (Swiss Federal Office of Transport, 2018), whereas the routes were obtained through Open Street Map (OpenStreetMap contributors, 2017). While we are aware of the ongoing debate about the respective merits of authoritative and volunteered data, we decided to use official data where available, and volunteered data where necessary.

For the moving windows over which the quantiles were to be calculated, we chose 130 seconds. This was big enough to allow multiple points within the windows for all but the largest temporal granularities and was in line with the orders of magnitude that can be found in the literature (Bolbol et al., 2012; Stenneth et al., 2011; Ellis et al., 2014). For the quantiles themselves we chose quartiles. The goal was to have a value representing the central tendency (Q2) and two that represented high (Q3) and low (Q1) values respectively. To avoid the detrimental impact of outliers we opted against taking the averages and extremal values sometimes found in the literature (Bohte & Maat, 2009; Gonzalez et al., 2008).

For all the numeric values (that all happen to be non-negative), we applied a log-transform to help less robust classifiers and used scaling to achieve zero mean and unit variance for all of them.

In addition, for a point we have added the quartiles of the above features for all points by the same user that were recorded within a certain radius and a temporal window. The reasoning behind this is that similar positions could mean similar labels, in which case having data from the past could contribute to averaging out errors incurred by the imprecise tracking.

3.7. Classification methods

As mentioned, we wanted to have representatives of the most common approaches to classifying modes of transport in our study: Pointwise classification, pointwise classification followed by some smoothing, segmentwise classification, and approaches integrating segmentation and classifying in one step.

Pointwise classification is straightforward and tries to find a mapping from the features of every individual point in a trip to the most likely mode of transport. The resulting stages are implicit and based on how many consecutive points share the same predicted label. As classifiers we used KNN, logistic regressions (LL), random forests (RF), and support vector machines (SVM). In addition and not fully compliant with the idea of a pointwise classifier, we used conditional random fields (CRF). All of those methods have been used with varying degrees of success to classify transport modes and for a more in-depth discussion on them we refer to Section 2.

For the optional smoothing of the inferred labels we have used different schemes. The first is a simple majority vote over a number of points that would correspond to two minutes if the points were regularly sampled (e.g. 4 for the case where the points are subsampled to 30 seconds). However, there were always at least 3 points, such that even for the temporally coarsest case there would always be real smoothing.

The second smoothing approach uses a Hidden Markov Model (HMM) on the predicted probabilities of the labels, following the ideas in Nitsche et al. (2014). We learn the HMM in a supervised fashion on the training data using the predicted class probabilities on the training data in combination with the true labels and applying the fitted HMM on the outcome of the predictions for the test data.

Finally, we reused the CRF idea from the pointwise classification as a post-classification smoother. The training procedure was the same as for the HMM smoother.

For the segmentwise classification we partitioned every trip at spatiotemporal points where the speed was below 1 km/h for 130 seconds. Such simple thresholds on speed and duration are quite common in the literature (Biljecki et al., 2013; Stopher et al., 2008; Chung & Shalaby, 2005) and there would conceivably be some accuracy to be gained by devising more sophisticated segmentation schemes.

To the segmentwise classifications we also added the two smoothing regimes described above. While the smoothing on the pointwise classification is mainly motivated by the elimination of stray labels, the main reason to also apply it on the labels for the segments is to avoid unreasonable combinations such as car-bike-car.

We also wanted a method that combines elements of segmentation and classification. We chose conditional random fields, as for the smoothing above. As the method encourages realistic sequences of labels, it has to receive special treatment when interpreting the results. In addition, despite not yet being well established, we added a representative of the deep learning family, deep recurrent neural networks (RNN) using bi-LSTM layers, specifically the network used by Zhao et al. (2019). We also implemented the idea of Simoncini et al. (2018) that focusses on feature extraction by means of dense feed-forward layers before the LSTM layers, but this yielded inferior results, which is why we do not discuss this approach further. Due to the significant run times of deep networks we skipped the cross-validation for these classifiers, as is common in the deep learning literature but still report the results from the test set comprised of trips from users not seen in training to allow for a fair comparison.

3.8. Evaluation of classified results

Rooted in the different applications for which transport mode detection is being used, there is a distinct lack of consensus as to how mode of transport detection should be evaluated and there exists no benchmark dataset on which all methods are evaluated. While we cannot alleviate the second problem here, we can try to give different evaluation metrics that allow accommodating different kinds of research questions.

One of the more popular metrics is the accuracy which counts the percentage of correct labels in the evaluation dataset (Ellis et al., 2014; Reddy et al., 2010; Semanjski et al., 2017). This measure is well suited if only the overall proportion of the different kinds of transport modes is important, e.g. in the context of location based services. It may, however, be somewhat problematic if the dataset is highly skewed, as the less frequent transport modes will tend to be under-represented in the labels.

An alternative is precision and recall by mode of transport, e.g. reported using a confusion matrix (Mäenpää et al., 2017). This has the advantage that if one mode is of particular interest, the error associated with it can be directly read off. However, confusion matrices can in general not be ordered and thus it is not possible to determine a "best" method. To have a single number for comparisons, Cohen's Kappa can be calculated to summarise the matrix (Bolbol et al., 2012; Huss et al., 2014).

Lastly if one is interested in "representative" trips, then the sequence is of particular interest and measures of differences between sequences must be used, such as the edit distance (Chen et al., 2005). This measure has been extended for information needs that go beyond the sequence as such (Prelipcean et al., 2016), but as we want to present how accurately the methods describe the sequences of transportation modes, we will stick to the edit distance.

3.9. Cross-validation

Typically, studies limit themselves to splitting the data into training and validation datasets and reporting the point estimates of the chosen evaluation criteria. However, to also gain insights on how strongly those estimates can vary, we perform a 10 fold cross-validation. This cross validation is done on the user level, i.e. splitting the users into 10 bins, 9 of which are used in training for every fold. Thus, we avoid producing results that are overly optimistic as a result of data of the same users being used in training and testing.

4. Results

We will first present the effect of the deterioration of spatial accuracy on the pointwise classifiers, followed by the results from the segmentwise classifiers (both without longitudinal data). Following that, we will present a list of results obtained by deviating from some of the choices we made for this study. We end the section by providing the confusion matrices.

While the complete results can be found in the Supporting Materials, we only include here the K-nearest neighbours (KNN), Random Forests (RF) and Conditional Random Fields (CRF) results. KNN is chosen as a benchmark that serves as a clear lower bound of what one would expect from a classifier. RF was kept as it produces the best results of the simple (i.e. truly pointwise) classifiers while sharing with the other classifiers the same qualitative behaviour. CRF as a classifier that inherently considers sequences of points can be expected to behave differently from the rest and therefore warrants discussion.

On the classifiers whose results are additionally included in the Supporting Materials, we only briefly summarise the overall impression: The logistic regression overall yielded results that lie somewhere between those obtained from KNN and RF. The runtime of the SVM algorithm used is above quadratic (Pedregosa et al., 2011), preventing it from being applied to the whole training set. Therefore it was only applied on a subset of the training data. The decrease between SVM and RF is smaller for the segmentwise classification than for pointwise, as the number of segments in the training set is significantly smaller than the number of points, meaning that the SVM classifier sees a larger proportion of the points. But even in the case of the segmentwise classification the results of SVM classification remained below those of the RF.

4.1. Pointwise classification

We start by describing the results from the non-distorted results shown in Figure 5. The numbers from the pointwise classifications are what could have been expected.

Non-trivial classifiers such as RF clearly outperform KNN, because (in the case of RF) they look at more relevant neighbourhoods in the feature space than simple spheres, as is the case in KNN. RF in turn gets dominated by CRF, which again is not too surprising, as CRFs can look at more than just the features of a single point to determine its class.

In terms of accuracy and Kappa, for the non-CRF classifiers, sparser temporal sampling seems to coincide with better results. Note that in a sparser sample, features such as velocity average over a longer time, resulting in more context information being available in the features of a single point. After smoothing, however, the results of different temporal granularities are comparable within a single classifier.

The edit distances of both KNN and RF are well above 1 (and at times above 10) and therefore indicate that those methods should not be used, if one is interested in the sequence of the transport modes. CRFs on the other hand have a markedly lower edit distance and the average of edits needed to each sequence is well below one.

The results described above, stemming from pure classification differ markedly from the results from pipelines that have a smoothing step after the classification. In particular, CRF no longer compares as favourably to the other classifiers. The HMM smoother is bad after KNN classification, but yields good results after a random forest classification. If a CRF smoother is used instead, the reverse is true.

Before turning to the results of the spatial distortion we would like to note the fact that the results reported here are below some of those found in the literature. For this we refer to the presentation on sensitivities later in the section.

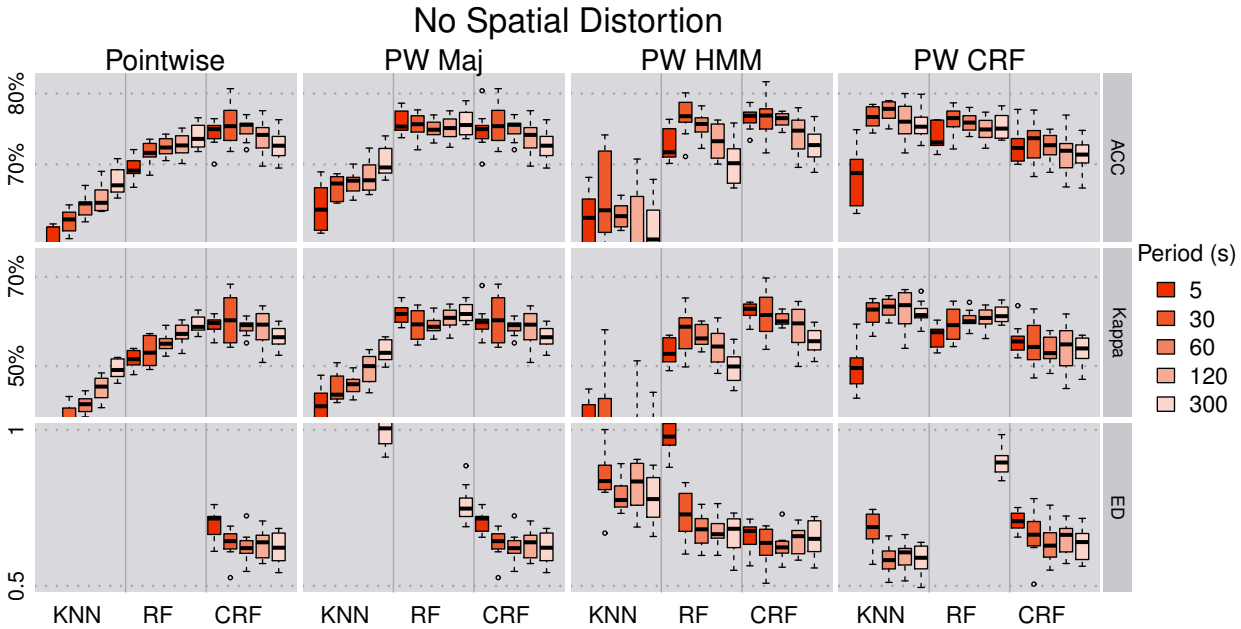


Figure 5: Results of the pointwise classification on the non-distorted points. Top row: Accuracy, Middle: Cohen's Kappa, Bottom: Edit Distance. The columns contain (from left to right): Pointwise estimator, smoothed with majority vote, smoothed with HMM, smoothed with CRF. Every plot contains the results grouped by classifier, ordered and coloured by temporal granularity. Every coloured box represents the ten values from the cross-validation.

If one applies a spatial distortion, the pure pointwise classification drops markedly, as seen when comparing Figures 5 and 6. The effect is stronger, the more temporally fine-grained the data are, as was to be expected. While the effect is somewhat mitigated if the pointwise results are smoothed, 5-second intervals still do not seem to be very useful for direct classification. The combination of a CRF initial classifier and a HMM smoother yields relatively stable results throughout the distortions.

4.2. Segmentwise classification

When looking at the results from the segmentwise classification – recall that the segmentation is neither learned nor known *a priori* but the result of thresholds on speed and time – there are several striking differences to the pointwise classification, as becomes evident when comparing Figures 5 and 7.

The two 'simple' classifiers KNN and RF benefit significantly from having features based on move segments, whereas the CRF classifier cannot benefit from them and now has results very similar to the very simple KNN classifier. RF, however, now obtains results that are an improvement over the best of those from the pointwise classification.

The second striking feature of the results is that any smoothing applied to the obtained results does no longer improve them. As a last difference to the pointwise case, the HMM smoother seems to produce distinctly worse results than the other two.

In stark contrast to the pointwise case, classifying segments seems a lot less sensitive to spatial distortions, as seen when comparing Figures 7 and 8. We believe that this finding echoes the one from the pointwise case, where the results obtained from temporally more coarsely resolved points were better. In both cases, the features are influenced less by distortions of the same order of magnitude, as they are based on points that are further apart. As particularly motorised segments (that abound in our data) can easily be rather long due to a lack of stops, the calculation of overall displacement and median speed are hardly affected by distortions, as they do not accumulate.

When comparing between Figures 8 and 6, i.e. the pointwise and the segmentwise classification on distorted data, it is striking that even with all the smoothing the results of pointwise classification are

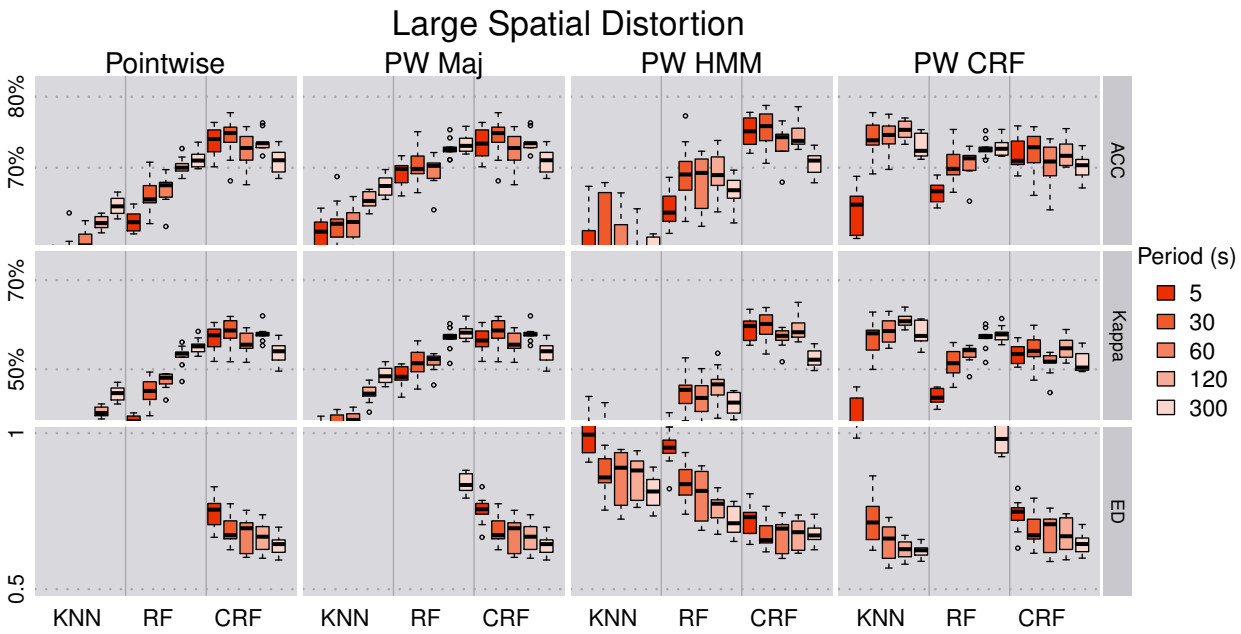


Figure 6: Results of the pointwise classification with strongly distorted points.

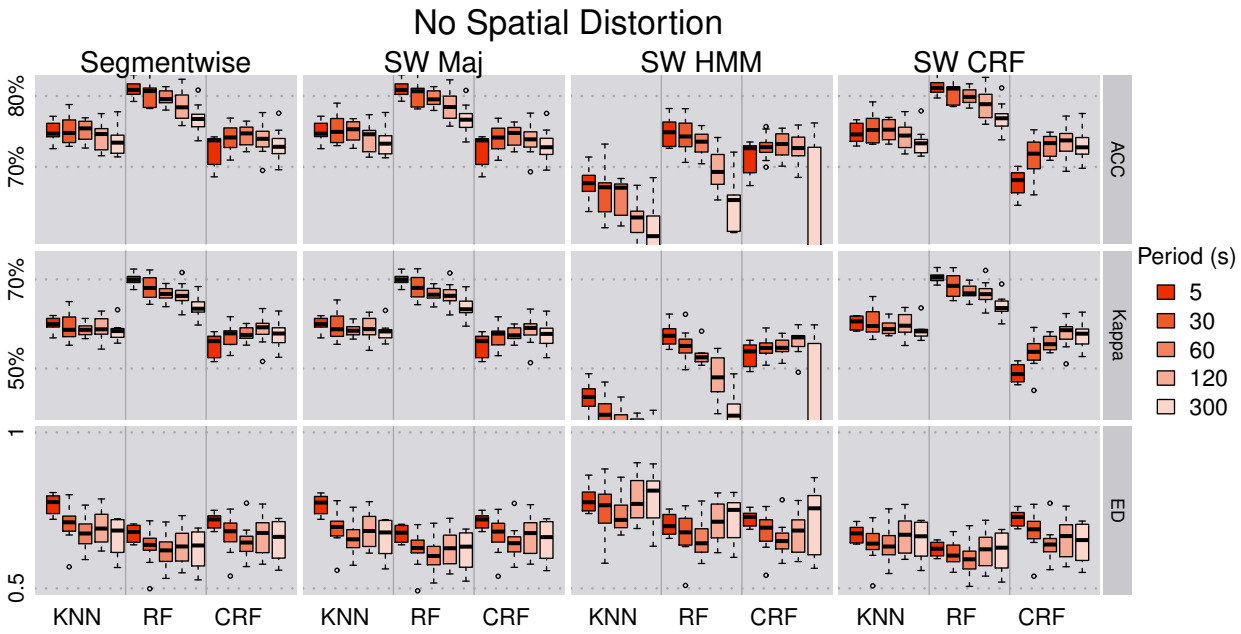


Figure 7: Results of the segmentwise classification on undistorted points.

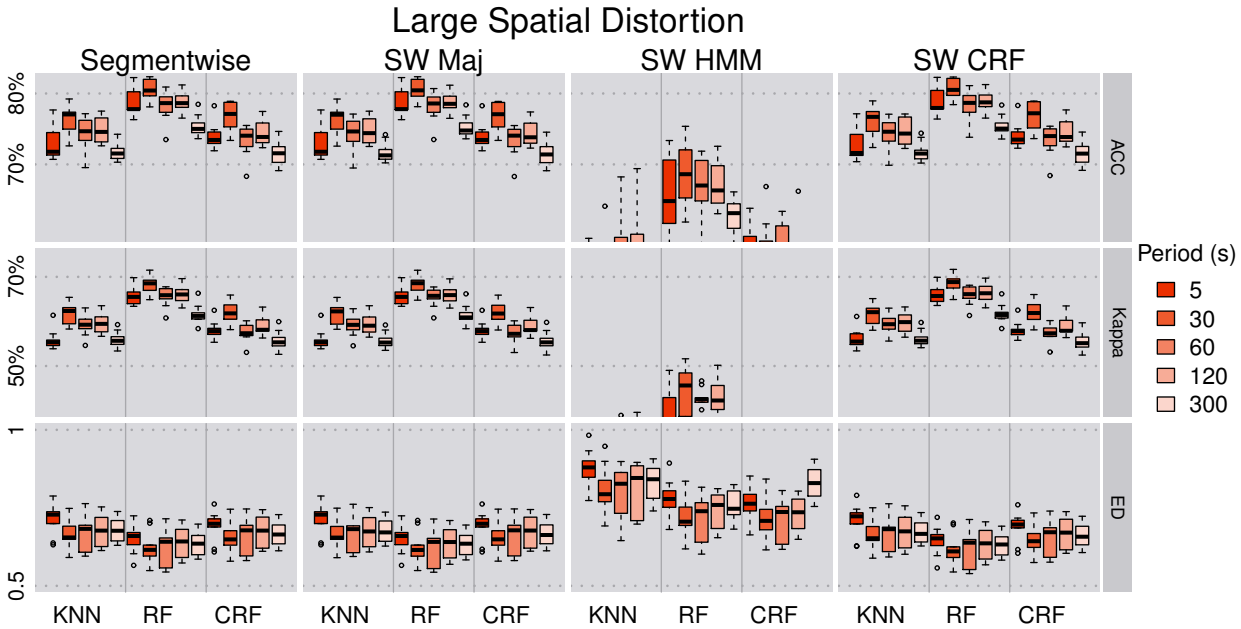


Figure 8: Results of the segmentwise classification with strongly distorted points.

simply not as good when classifying in a pointwise fashion than when classifying by segments.

Lastly, we present the evolution of the accuracies over the course of the deteriorations for three representative cases in Figure 9. While only three combinations amongst those given in Figures 5 – 8 are given, the actual deterioration is more clearly visible. Also, since Figure 9 contains non-CV figures, the results from the deep neural networks can also be shown.

4.3. Including longitudinal information

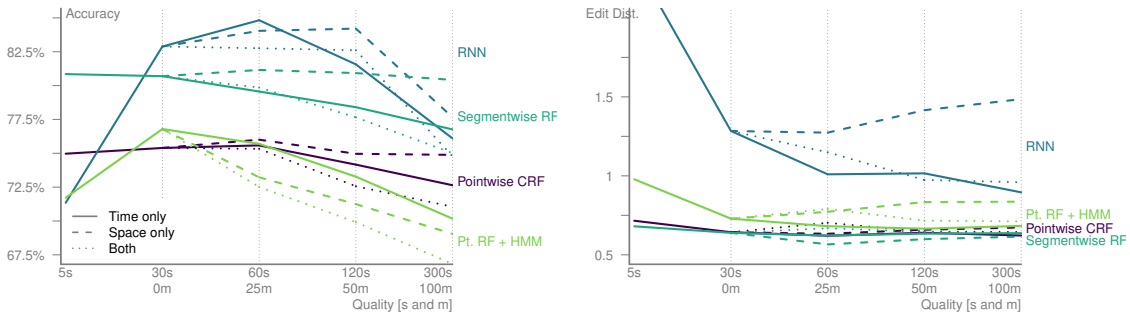
Including the longitudinal data does not change the picture dramatically. While there are some positive effects for the pointwise and smoothed classifications, especially for the noisy data the effect is smaller or even reversed in the case of segmentwise classification, where the best results are achieved. The detailed results can be found in the Supporting Materials, but we refrain from giving a detailed description here.

4.4. Sensitivities

A complete run through all the possible combinations of choices for all the parameters would not be possible due to the combinatorial explosion of different cases. However, to get some ideas about how certain choices could affect the results we performed a sensitivity analysis for a few select parameters that we held constant in the main experiment. The results of these sensitivities can be seen in Table 5, where the baseline corresponds to the 30-second and non-smoothed results from the RF classifier in Figure 7. The classifier, smoother, temporal granularity, and spatial uncertainty were held constant for all the comparisons.

The sensitivities tested were the following:

- *No GIS Information* removes all GIS dependent features from the data. This should allow for a cost-benefit consideration of taking the effort of adding this type of information when designing a system for passive tracking.
- Generous segmentation corresponds to an alternative segmentation scheme that creates more segments, namely whenever the speed falls below 10 km/h (for any length of time).
- CV by trip does the cross-validation treating every trip as equal instead of cross-validating by user.



(a) Accuracy

(b) Edit Distance

Figure 9: Decreases in accuracy and edit distance over deterioration for three representative approaches to mode detection plus the deep learning alternative. On the x-axis, the index of the spatial and/or temporal quality deterioration amongst the ordered chosen value is given. Due to the limited space, only one example of each of the three classes of classifiers tested here are given, and only for one measure: Pointwise plus smoothing, segmentwise, pointwise CRF and the deep RNN. The curves for the other measures and for other classification approaches look qualitatively similar. For the time only curves, the spatial uncertainty was set to zero, whereas for the space only curves, the temporal granularity was set to the 30 seconds recommended by Bolbol & Cheng (2010).

- GT segmentation, in addition to cross-validating by trips uses the ground truth segments for classification, which simplifies the problem, is frequently used in the literature and leads to good results.

This reiterates that even given the same classification method the results can differ significantly and thus the method should not be viewed in isolation, but always together with the problem statement and the data.

4.5. Confusion matrices

Lastly we would like to present the confusion matrices we obtained. Based on the results we saw for the different temporal windows, we show the matrices at the 30 second temporal granularity and the RF classifier based on segments – both inferred and ground truth.

We see in Table 6 for the undistorted data that the modes that are by far the most common in our dataset (car and train) get classified correctly most of the time, with a recall of about 90% each. This despite the fact that the skewness in the labels was accounted for when training the classifiers. However, the same results cannot be obtained for the modes that are less common. Particularly the local public transport modes *Bus* and *Tram* get mistaken surprisingly often for Cars or Trains.

Table 7 reveals that while the most common modes are hardly affected at all by the spatial distortion, the already quite poorly classified modes suffer particularly strongly.

Classification based on ground truth segmentation is unsurprisingly much better than if it is based on inferred segments, as seen in Table 8. In particular the walking segments benefit greatly. While Trams now are decently discovered, buses and bikes still suffer from poor recall values, even though the precision has improved significantly.

In terms of spatial distortion, the deterioration is less dramatic than for classification on inferred segments, as can be seen in Table 9. While it is still the modes with poor recall that suffer the most, the decline is smaller than before.

5. Discussion

5.1. Overall results

On the most important question, concerning what quality the data from passive tracking would need to deliver in order to allow traffic mode detection, we can observe the following:

In terms of temporal granularity, sampling at too high a frequency will not benefit the classification results and on the contrary even deteriorate them in pointwise approaches, particularly those that have built-in smoothing such as CRF rather than a combined classification and smoothing. In segmentwise approaches, having a very fine temporal granularity does not help much, but does not deteriorate the results that much.

Furthermore, there is no strong interaction with spatial accuracy and the above holds for all tested values. On the one hand, this confirms findings from the literature in the pointwise case (Bolbol & Cheng, 2010) that claim an ideal sampling rate in the order of magnitude of about a minute, but on the other hand it generalises them to the case where spatial accuracy of the measurements cannot be treated as a given. In the segmentwise case on the other hand where features are calculated over longer periods, the adverse effect of too fine grained sampling could not be observed.

In terms of spatial distortions, the picture is less clear. The traffic modes that made up the bulk of the data and that were well classified in the absence of spatial distortions continued to be correctly identified most of the time even for the largest spatial distortion that we tested, as evidenced by Figure 9, especially for those approaches that have a good classification to begin with, as seen in Table 7. However, the other modes, that already were poorly identified in the base case, suffered considerably under spatial distortions. We believe that some of the poor results for local public transport can be attributed to the labelling scheme that did not allow users to insert missing trip legs. The fact that many bus/tram trips comprise the access/egress walking trip legs as well means that the classifiers cannot reliably learn that bus/tram legs start at corresponding stations. In addition, as slow segments carry a local public transport label, the classifiers can no longer reliably learn that slow speeds are indicative of walking segments. The breadth of users targeted in the original data collection campaign thus came at a rather significant cost in classification quality. There is a notable difference between accuracy and edit distance. While many of the good classifiers can cope fairly well with relatively large spatial distortions if accuracy is measured, spatial distortion affects edit distance negatively already at moderate levels for those classifiers.

On all spatial distortions that we tested, there was no complete breakdown of the methods on the bulk of the data, i.e. on the *car* and *train* modes. Rather there was a steady decline from the baseline. This means that there is no clear minimum uncertainty (in the range we tested) beyond which detection becomes completely unfeasible. But clearly, the more accurate the data, the better the results, particularly in pointwise classification. This is in contrast to the temporal granularity, where too much detail could be detrimental.

In terms of the methods compared, it became clear that overall, the best approach seems to be to apply a decent segmentation and classify based on segments. This confirms common sense expectations that adding semantically meaningful context variables help classification. Smoothing is not necessary when segmenting first, but absolutely necessary if the classification happens in a pointwise fashion. In terms of classifiers, Random Forests had stable and qualitatively appealing results. While the deep recurrent neural networks provided superior results in terms of accuracy, they were somewhat lacking in performance when it comes to edit distance, where they performed significantly worse than either of the more standard approaches. Thus, when using the early network architectures we tested, there seems to be a trade-off and for every use case it has to be decided, whether the reduced edit-distance is acceptable. Alternatively, network architectures such as sequence to sequence schemes could be considered to obtain more reasonable label sequences.

5.2. Sensitivities

The results of the sensitivity analysis conformed to expectations, at least qualitatively.

GIS information does contribute to the classification, although not quite to the expected degree. If one looks at the feature importance for tram and bus (not included here for brevity), GIS features rank among the most important predictors. Interestingly this is not the case for train, where the instantaneous (calculated) speed at the end of a segment as well as overall displacement are the top features. With very few GNSS fixes inside trains, the beginning and end comprise more or less the trip as a whole, as does the displacement measured between them, since any stops in the middle cannot cause a segment break, leading to larger segments than are observed for other modes. Thus, for trains, GIS information may not be adding

much in cases such as ours, where there are few or no valid GNSS fixes. Thus, as the data was skewed away from the transport modes where GIS was helpful, the overall contribution was limited.

With respect to segmentation thresholds, having thresholds that result in more segments, as the one shown in the results, can lead to oversegmentation and hence to high edit distances. While smoothing can remedy some of this deterioration (as shown in the Supporting Materials only), it does not lead to results that beat that of the baseline.

Using a cross-validation scheme that cross-validates by trip instead of by user, the results get slightly better, but mostly the standard deviation decreases substantially. The folds used in the trip based scheme do not fundamentally differ from one another, since all folds contain trips from all users, leading to very stable, but overly optimistic results, underestimating the uncertainty in the quality measures when generalising to people that did not contribute to the training data.

The results from using the ground-truth segmentation underline the importance of good segmentation. The closer the segments get to stages, the better the expected results. It also shows that the good results reported on pre-segmented trips should not be used to form expectations about the classification accuracies in situations when the trip legs are not given.

5.3. Confusion matrices

Overall, even with features limited to those available to passive tracking schemes, the overall accuracies were in the range expected from the literature. However, some modes were quite poorly identified.

For the Walk label, this can partly be explained by labelling issues discussed in Section 3.1: There were plenty of very slow segments in bus stages during training and therefore, while all seen walking stages are slow, not all slow segments that should have done so belonged to walking stages.

Regardless of the segmentation used, bicycles were not that easy to detect. They seem to take some place between walking and cars. This appears plausible, as a bicycle leg can look almost as one on foot if it is uphill, or can be nearly indistinguishable from a car in city traffic, if it shares the same restrictions in terms of traffic lights, stop signs, or similar. While this distinction is easier when accelerometers are available, distinguishing the three modes is much harder based on GNSS alone.

The classifications based on the ground truth stages can help to shed at least some light on the effect of the less than perfect labels. To be clear, the effect is confounded with the fact that the problem of labelling stages rather than segments is easier, but we believe there are some pointers nonetheless. Most striking is the increase in quality for the walk stages. The deterioration of the results that derives from the deterioration in spatial accuracy is still clearly visible, but less extreme than for the results on the move segments.

As slow segments are no longer seen in isolation (as walking stages were often merged into stages of other modes), significantly more of the slow segments actually reflect walking stages, which makes the stages labelled *Walk* significantly more separable in the feature space.

Again, as for the inferred case, Trams are more easily identifiable than buses, due to the fact that there are fewer cities in which there are trams in the first place, making the GIS information more useful here (as reflected in the higher feature importance). The buses are still hard to separate from cars, but this does not come entirely surprising, since they do share similar characteristics.

6. Conclusion and outlook

In this work we have applied commonly used methods for the classification of traffic modes to information that could be available from passively sensing mobile phone data through the mobile phone network for various levels of temporal granularity and spatial uncertainty. This may help focus the priorities for data quality improvements by actors using passive tracking and aiming at performing transport mode detection, such as telephone companies. We have used realistic data from over 130 users collected over half a year, which was annotated by those users.

The answer to the question what levels of spatial accuracy and temporal granularity are required to perform mode detection depend strongly on the scheme that is chosen for the classification.

In cases where the only option is the pointwise classification – for example in real-time classification – sampling intervals between 30 seconds and a minute seem optimal. Values both below and above that window will decrease the quality of the classification. In terms of spacial accuracies, a higher accuracy is always preferable. The reduction in classification quality seems to be approximately linear in the standard deviation of the sampling error.

In cases where the whole trajectories are available (segmentwise or CRF), contextual information mitigates the effect of the deterioration of quality. If viewed in isolation, spatial uncertainty alone is much less detrimental than temporal sparsity.

This clarifies the priorities with which improvements in passive data collection should be addressed: First, efforts should be made to bring the sampling rates to the orders of magnitude between 30 seconds and a minute. As the effect of spatial accuracy is less of a concern, particularly when temporal granularity is high, it should be addressed with lower priority.

For this study, we had at our disposal a very wide range of people contributing their data, distributed across a wide range of geographical situations over a long time. This came at the cost of a reduced interaction when labelling the data, leading to some fused stages. It is not possible to tell how strongly this problem affected the data quality, but the results shown here can be seen as a lower bound of what is possible. Confidently separating more modes may be possible with more accurate labels.

As for the most successful classification strategy, irrespective of quality, the best results were obtained by segmenting a trip into meaningful parts and classifying based on segmentwise features. Random forests have yielded the best overall results in this setting. While deep recurrent neural networks can be an alternative, especially when the focus lies on accuracy as a measure, they did not produce sensible sequences as evidenced by the higher edit distance in Figure 9.

Future work could expand on the classifiers and segmentations used in this paper. It could in particular also try to identify situations in which methods perform particularly well or badly to further the understanding of *why* they do or do not work. This holds especially true for deep neural networks whose inner workings are notoriously hard to interpret.

We believe that traffic mode detection based on passively sensed data is not yet satisfactorily solved. In particular there are two areas where we see need for additional work. The first is finding segmentations whose resulting segments are closer to stages. As indicated by the leap in classification quality observed when using ground truth segments, there still seems to be untapped potential.

In addition, to best allocate research resources in the future it might be beneficial to identify the properties of a training data set that are most important to assure a low generalisation error. Properties that come to mind are the geographies that need to be captured in training, the diversity amongst the population recording the training data when compared to the target population, or the balance of the modes in training. If, for example, an approach generalises well from one group of people to another then it may not be necessary to recruit a large number of people, and having a few dedicated researchers ensure more important properties, e.g. mode balance, will be a better allocation of resources.

Lastly, having a standard dataset which researchers have access to and agree on to test their methods on would go a long way to alleviate the problem of incomparable results. Unfortunately in the collection of the proprietary dataset used in this research the participants were not asked to consent to their tracks (and by extension the locations of their homes and work places) being disclosed so we cannot publish the dataset.

References

Axhausen, K., Zimmermann, A., Schönfelder, S., Rindsfüser, G., & Haupt, T. (2002). Observing the rhythmns of daily life: A six week travel diary. *Transportation*, *29*, 95–124.

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

Bantis, T., & Haworth, J. (2017). Who you are is how you travel: A framework for transportation mode detection using individual and environmental characteristics. *Transportation Research Part C: Emerging Technologies*, *80*, 286–309.

Becker, H., Loder, A., Schmid, B., Jonietz, D., Bucher, D., Raubal, M., & Axhausen, K. W. (2018). Usage patterns and impacts of a mobility flat rate traced with a Smartphone App. In *15th International Conference on Travel Behaviour Research, (IATBR2018), Santa Barbara*. URL: `https://www.research-collection.ethz.ch/handle/20.500.11850/312525`.

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

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*, 285–297.

Bolbol, A., & Cheng, T. (2010). GPS Data Collection Setting For Pedestrian Activity Modelling. *Proceedings of the GIS Research UK 18th Annual Conference GISRUK 2010*, (pp. 337 – 344).

Bolbol, A., Cheng, T., Tsapakis, I., & Haworth, J. (2012). Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. *Computers, Environment and Urban Systems*, *36*, 526–537.

Bricka, S., & Bhat, C. (2006). Comparative analysis of global positioning system-based and travel survey-based data. *Transportation Research Record: Journal of the Transportation Research Board*, (pp. 9–20).

Bricka, S., Zmud, J., Wolf, J., & Freedman, J. (2009). Household travel surveys with GPS: An experiment. *Transportation Research Record: Journal of the Transportation Research Board*, (pp. 51–56).

Burkhard, O., Ahas, R., Saluveer, E., & Weibel, R. (2017). Extracting regular mobility patterns from sparse CDR data without a priori assumptions. *Journal of Location Based Services*, .

Chen, C., Gong, H., Lawson, C., & Bialostozky, E. (2010). Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. *Transportation Research Part A: Policy and Practice*, *44*, 830–840.

Chen, L., Özsu, M. T., & Oria, V. (2005). Robust and fast similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data - SIGMOD '05* (p. 491).

Chin, K., Huang, H., Horn, C., Kasanicky, I., & Weibel, R. (2019). Computers , Environment and Urban Systems Inferring fine-grained transport modes from mobile phone cellular signaling data. *Computers, Environment and Urban Systems*, *77*. URL: `https://doi.org/10.1016/j.compenvurbsys.2019.101348`.

Chung, E. H., & Shalaby, A. (2005). A trip reconstruction tool for GPS-based personal travel surveys. *Transportation Planning and Technology*, *28*, 381–401.

Csáji, B. C., Browet, A., Traag, V. a., Delvenne, J. C., Huens, E., Van Dooren, P., Smoreda, Z., & Blondel, V. D. (2013). Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, *392*, 1459–1473.

Dammann, A., Raulefs, R., & Zhang, S. (2015). On prospects of positioning in 5G. In *Communication Workshop (ICCW), 2015 IEEE International Conference on* (pp. 1207–1213). IEEE.

Eftekhari, H. R., & Ghatee, M. (2016). An inference engine for smartphones to preprocess data and detect stationary and transportation modes. *Transportation Research Part C*, *69*, 313–327.

Ellis, K., Godbole, S., Marshall, S., Lanckriet, G., Staudenmayer, J., & Kerr, J. (2014). Identifying Active Travel Behaviors in Challenging Environments Using GPS, Accelerometers, and Machine Learning Algorithms. *Frontiers in Public Health*, *2*, 1–8.

Feng, T., & Timmermans, H. J. (2016). Comparison of advanced imputation algorithms for detection of transportation mode and activity episode using GPS data. *Transportation Planning and Technology*, *39*, 180–194.

Furletti, B., Cintia, P., Renso, C., & Spinsanti, L. (2013). Inferring human activities from gps tracks, . (p. 5).

Gong, H., Chen, C., Bialostozky, E., & Lawson, C. T. (2012). A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems*, *36*, 131–139.

Gonzalez, P. A., Weinstein, J. S., Barbeau, S. J., Labrador, M. A., Winters, P. L., Georggi, N. L., & Perez, R. A. (2008). Automating Mode Detection Using Neural Networks and Assisted GPS Data Collected Using GPS-Enabled Mobile Phones. *15th World Congress on Intelligent Transport Systems*, (p. 12p).

Huang, H., Cheng, Y., & Weibel, R. (2019). Transport mode detection based on mobile phone network data : A systematic review. *Transportation Research Part C*, *101*, 297–312.

Huss, A., Beekhuizen, J., Kromhout, H., & Vermeulen, R. (2014). Using GPS-derived speed patterns for recognition of transport modes in adults. *International Journal of Health Geographics*, *13*, 40.

Jahangiri, A., & Rakha, H. A. (2015). Applying Machine Learning Techniques to Transportation Mode Recognition Using Mobile Phone Sensor Data. *Ieee Transactions on Intelligent Transportation Systems*, *16*, 2406–2417.

Jiang, S., Ferreira, J., & Gonzalez, M. C. (2017). Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore. *IEEE Transactions on Big Data*, *3*, 208–219.

Koivisto, M., Hakkarainen, A., Costa, M., Talvitie, J., Heiska, K., Leppanen, K., & Valkama, M. (2017). Continuous high-accuracy radio positioning of cars in ultra-dense 5G networks. *2017 13th International Wireless Communications and Mobile Computing Conference, IWCMC 2017*, (pp. 115–120).

Leber, J. (2013). How Wireless Carriers Are Monetizing Your Movements. URL: `https://www.technologyreview.com/s/513016/how-wireless-carriers-are-monetizing-your-movements/`.

Leuzzi, F., Del Signore, E., & Ferranti, R. (2017). Towards a Pervasive and Predictive Traffic Police. In *Italian Conference for the Traffic Police* (pp. 19–35). Springer.

Lin, Z., Yin, M., Feygin, S., Sheehan, M., Paiment, J.-F., & Pozdnoukhov, A. (2017). Deep Generative Models of Urban Mobility. *In Review*, .

Linnap, M., & Rice, A. (2014). Managed Participatory Sensing with YouSense. *Journal of Urban Technology*, *21*, 9–26.

Mäenpää, H., Lobov, A., & Martinez Lastra, J. L. (2017). Travel mode estimation for multi-modal journey planner. *Transportation Research Part C: Emerging Technologies*, *82*, 273–289.

Marra, A. D., Becker, H., Axhausen, K. W., & Corman, F. (2019). Developing a passive GPS tracking system to study long-term travel. *Transportation Research Part C*, *104*, 348–368.

Moiseeva, A., & Timmermans, H. (2010). Imputing relevant information from multi-day GPS tracers for retail planning and management using data fusion and context-sensitive learning. *Journal of Retailing and Consumer Services*, *17*, 189–199.

Müller, P., Del Peral-Rosado, J., Piché, R., & Seco-Granados, G. (2016). Statistical Trilateration With Skew-t Distributed Errors in LTE Networks. *IEEE Transactions on Wireless Communications*, *15*, 7114–7127.

Musolesi, M. (2014). Big mobile data mining: Good or evil? *IEEE Internet Computing*, *18*, 78–81.

Ni, L., Wang, X. C., & Chen, X. M. (2018). A spatial econometric model for travel flow analysis and real-world applications with massive mobile phone data. *Transportation Research Part C: Emerging Technologies*, *86*, 510–526.

Nitsche, P., Widhalm, P., Breuss, S., Brändle, N., & Maurer, P. (2014). Supporting large-scale travel surveys with smartphones A practical approach. *Transportation Research Part C*, *43*, 212–221.

Nitsche, P., Widhalm, P., Breuss, S., & Maurer, P. (2012). A strategy on how to utilize smartphones for automatically reconstructing trips in travel surveys. *Procedia - Social and Behavioral Sciences*, *48*, 1033–1046.

Nyhan, M., Sobolevsky, S., Kang, C., Robinson, P., Corti, A., Szell, M., Streets, D., Lu, Z., Britter, R., Barrett, S. R. H., & Others (2016). Predicting vehicular emissions in high spatial resolution using pervasively measured transportation data and microscopic emissions model. *Atmospheric Environment*, *140*, 352–363.

OpenStreetMap contributors (2017). Country extract from https://download.geofabrik.de . URL: `https://www.openstreetmap.org`.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pereira, F., Carrion, C., Zhao, F., Cottrill, C. D., Zegras, C., & Ben-Akiva, M. E. (2013). The Future Mobility Survey: overview and preliminary evaluation. *Proceedings of the Eastern Asia Society for Transporation Studies*, *9*.

Prelipcean, A. C., Gidofalvi, G., & Susilo, Y. O. (2016). Measures of transport mode segmentation of trajectories. *International Journal of Geographical Information Science*, *8816*, 1–22.

Prelipcean, A. C., Gidófalvi, G., & Susilo, Y. O. (2017). Transportation mode detection an in-depth review of applicability and reliability applicability and reliability. *Transport Reviews*, *37*, 442–464.

Reddy, S., Burke, J., Estrin, D., Hansen, M., & Srivastava, M. (2008). Determining transportation mode on mobile phones. *Proceedings - International Symposium on Wearable Computers, ISWC*, (pp. 25–28).

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., & Srivastava, M. (2010). Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks*, *6*, 1–27.

Rinzivillo, S., Gabrielli, L., Nanni, M., Pappalardo, L., Pedreschi, D., & Giannotti, F. (2014). The Purpose of Motion : Learning Activities from Individual Mobility Networks. *International Conference on Data Science and Advanced Analytics (DSAA14)*, .

Rollier, R. (2015). From big data to smart data, traffic optimization using mobile network traces. In *Smart City Suisse*.

Saeb, S., Lattie, E. G., Schueller, S. M., Kording, K. P., & Mohr, D. C. (2016). The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ*, *4*, e2537.

Sauerländer-Biebl, A., Brockfeld, E., Suske, D., & Melde, E. (2017). Evaluation of a transport mode detection using fuzzy rules. *Transportation Research Procedia*, *25*, 591–602.

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

Semanjski, I., Gautama, S., Ahas, R., & Witlox, F. (2017). Spatial context mining approach for transport mode recognition from mobile sensed big data. *Computers, Environment and Urban Systems*, *66*, 38—-52.

Shafique, M. A., & Hato, E. (2015). Use of acceleration data for transportation mode prediction. *Transportation*, *42*, 163–188.

Shah, R. C., Wan, C.-y., Lu, H., & Nachman, L. (2014). Classifying the mode of transportation on mobile phones using GIS information. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '14 Adjunct*, (pp. 225–229).

Shen, L., & Stopher, P. R. (2014). Review of GPS Travel Survey and GPS Data-Processing Methods. *Transport Reviews*, *34*, 316–334.

Simoncini, M., Taccari, L., Sambo, F., Bravi, L., Salti, S., & Lori, A. (2018). Vehicle classi fi cation from low-frequency GPS data with recurrent neural networks. *Transportation Research Part C*, *91*, 176–191.

Stenneth, L., Wolfson, O., Yu, P. S., & Xu, B. (2011). Transportation mode detection using mobile phones and GIS information. *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, (p. 54).

Stopher, P., FitzGerald, C., & Zhang, J. (2008). Search for a global positioning system device to measure person travel. *Transportation Research Part C: Emerging Technologies*, *16*, 350–369.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *NIPS* (pp. 3104–3112). `arXiv:1409.3215`.

Swiss Federal Office of Transport (2018). Open Data Platform Swiss Public Transport. URL: `https://opentransportdata.swiss/`.

Swisscom AG (2018). *Annual Report*. Technical Report.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *CoRR*, *abs/1706.0*. `arXiv:1706.03762`.

Xiao, G., Juan, Z., & Zhang, C. (2015). Travel mode detection based on GPS track data and Bayesian networks. *Computers, Environment and Urban Systems*, *54*, 14–22.

Zhang, L., Dalyot, S., Eggert, D., & Sester, M. (2012). Multi-Stage Approach To Travel-Mode Segmentation and Classification of Gps Traces. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XXXVIII-4/*, 87–93.

Zhao, H., Hou, C., Alrobassy, H., & Zeng, X. (2019). Recognition of Transportation State by Smartphone Sensors Using Deep

Bi-LSTM Neural Network. *Journal of Computer Networks and Communication*, .

Zheng, Y., Liu, L., Wang, L., & Xie, X. (2008). Learning transportation mode from raw gps data for geographic applications on the web. *Proceeding of the 17th international conference on World Wide Web - WWW '08*, (p. 247).

Zhu, X., Li, J., Liu, Z., Wang, S., & Yang, F. (2016). Learning transportation annotated mobility profiles from GPS data for context-aware mobile services. *Proceedings - 2016 IEEE International Conference on Services Computing, SCC 2016*, (pp. 475–482).

Zolliker, M., Rollier, R., & Bosshard, A. (2015). Co-creation eines Smart-Data-Werkzeuges zur Verkehrsmessung. In *4. nationale Smart City Tagung*.

Classified entity	Special features	Participants and Geographic spread	Training and Testing	Study
Point	Personal Preferences	2 Users, London, up to seven days	Training & testing on same users	Bantis & Haworth (2017)
Unimodal trips	Instantaneous speed	1104 Homeowners in three cities for one week	No training/testing split mentioned	Bohte & Maat (2009)
Point	Instantaneous speed	81 Users, 2 Weeks in London	No training/testing split mentioned	Bolbol et al. (2012)
Segments		49 participants, New York, up to 5 days	No separate training and Testing	Chen et al. (2010)
Unimodal stages		99 Trips by 2 Volunteers and 48 days by 9 Volunteers in Graz and Vienna	Trip-wise cross validation	Chin et al. (2019)
1 Minute Intervals	Signal to noise ratio of GPS signal	Controlled movement protocol	150h, 2 research assistants	Ellis et al. (2014)
Point	Vehicle ownership	8 participants, up to 8 weeks, two cities	Overlap of moving windows between training and testing data	Feng & Timmermans (2016)
Segment		49 single days, students and employees	No split of training and test data	Gong et al. (2012)
Unimodal trips	Instantaneous speed	114 Trips in Tampa, Florida	Cross-Validation by Trip	Gonzalez et al. (2008)
Unimodal trips	Instantaneous speed	2 Commutes of 12 university employees	Cross-Validation by person	Huss et al. (2014)
Unimodal stages		182 GeoLife Participants and other sources, up to multiple months, different countries	No training/testing split mentioned	Mäenpää et al. (2017)
Stage	Realtime Public transport data	666 Users in Basel and Zürich	N/A	Marra et al. (2019)
N/A		Up to 250 minutes from four users	Vehicle Ownership	Moiseeva & Timmermans (2010)
5 Second Intervals	Accelerometry	266 hours, collected by 14 participants	Data on same	Nitsche et al. (2012)
Short windows up to 120s	Accelerometry	355h by 15 volunteers	Cross-validation by time window	Nitsche et al. (2014)
Unimodal segments		102 days by 26 users	No training / testing split mentioned	Pereira et al. (2013)
1 second windows	WiFi	20h by 16 individuals	Cross-validation by time window	Reddy et al. (2010)
1 min windows	Acc. & Magnetometer	17 hours by 9 users	Validation on Training data	Eftekhari & Ghatee (2016)
Pre-Segmented Trips		8303 Users, 4 months, one city	Training & testing on same users & testing on same users	Semanjski et al. (2017)
Overlapping 10 minute windows		47 Participants, 2 months, 3 cities in Japan	Training / testing split by window	Shafique & Hato (2015)
5 second window	GIS	50hours, 15 Individuaals, San Franciso and Portland	No training / testing split mentioned	Shah et al. (2014)
30 second window	Real time GIS information	6 individuals, 3 weeks	10-fold CV by window	Stenneth et al. (2011)
Stages		1246 person days, 202 participants in Shanghai		Xiao et al. (2015)
Stages		125 trips in Hannover	Training / testing split by stage	Zhang et al. (2012)
Stages	GIS	182 Users, mostly in China	No training / testing split mentioned	Zhu et al. (2016)
Multiple for comparison		138 participants spread over the country, 6 months	Cross-Validation by person	This study

Table 1: Selection of related work illustrating the diversity of ways in which the problem mode detection is understood in the literature in terms of problem definition and generalisability that can be expected.

Mode Sequence	Count
Car	38,382
Walk	22,251
Car, Walk	3,812
Walk, Car	2,909
Bicycle	2,842
Train	2,098
Train, Walk	1,439
Walk, Train, Walk	1,025
Walk, Train	987
Bus	699
Total	90,515

Table 2: Most common label sequences on trips. Even among the most common sequences there are some that are not what one would expect from theory (e.g. Train without walking stage leading up to it).

Mode	Number of points
Car	6,091,407
Train	3,050,785
Walk	1,984,240
Bike	511,906
Bus	223,977
Tram	116,760

Table 3: Number of ground truth labels (on a 5 second basis) for the modes under study. The distribution is very uneven, but reflects the behaviour of the population under study over the time span of the data collection.

t	Decision	Reason
16:00:00	Keep	First in Interval 16:00-16:05
16:01:00	Drop	Second in Interval 16:00-16:05
16:14:59	Keep	First in Interval 16:10-16:15
16:15:01	Keep	First in Interval 16:15-16:20

Table 4: Illustration of the temporal subsampling. The actual data was collected at 1 Hz. The fictive timestamps given here are simply used to demonstrate the subsampling method.

Sensitivity	Accuracy	Acc. SD	Kappa	Kappa SD	Edit Dist	ED SD
Baseline	80.43	1.47	0.69	0.69	0.62	0.05
No GIS Information	79.57	1.57	0.67	0.67	0.60	0.05
Generous segmentation	77.21	1.16	0.64	0.64	2.03	0.13
CV by trip	80.46	0.35	0.69	0.69	0.61	0.01
GT segmentation	90.03	0.42	0.84	0.84	0.24	0.01

Table 5: Median values and standard deviations for the cross-validated quality measures. The temporal granularity was always 30 seconds and the spatial distortion was not present to produce the results. See Section 4.4 for an explanation of the sensitivities.

	Car	Bike	Walk	Train	Tram	Bus	Precision
Car	90%	45%	20%	10%	12%	57%	83%
Bike	0%	32%	1%	0%	2%	0%	86%
Walk	3%	12%	57%	2%	19%	10%	77%
Train	7%	11%	21%	87%	23%	20%	76%
Tram	0%	0%	1%	0%	45%	1%	53%
Bus	0%	0%	0%	0%	0%	12%	56%
Recall	90%	32%	57%	87%	45%	12%	

Table 6: Confusion matrix with no spatial distortion for the combined segmentation and classification problem.

	Car	Bike	Walk	Train	Tram	Bus	Precision
Car	89%	42%	20%	7%	14%	50%	83%
Bike	0%	19%	0%	0%	1%	0%	87%
Walk	3%	20%	55%	1%	25%	11%	75%
Train	8%	19%	24%	91%	31%	35%	75%
Tram	0%	0%	1%	0%	29%	0%	57%
Bus	0%	0%	0%	0%	0%	3%	48%
Recall	89%	19%	55%	91%	29%	3%	

Table 7: Confusion matrix with large spatial distortion for the combined segmentation and classification problem.

	Car	Bike	Walk	Train	Tram	Bus	Precision
Car	95%	48%	8%	6%	5%	52%	90%
Bike	0%	31%	1%	0%	3%	0%	86%
Walk	2%	17%	90%	1%	14%	8%	87%
Train	3%	3%	0%	94%	2%	1%	95%
Tram	0%	0%	0%	0%	76%	1%	78%
Bus	0%	2%	0%	0%	0%	38%	75%
Recall	95%	31%	90%	94%	76%	38%	

Table 8: Confusion matrix with no spatial distortion for the pure classification problem.

	Car	Bike	Walk	Train	Tram	Bus	Precision
Car	94%	50%	8%	7%	8%	53%	86%
Bike	0%	25%	0%	0%	4%	0%	88%
Walk	3%	22%	91%	1%	25%	18%	83%
Train	3%	2%	0%	92%	2%	2%	95%
Tram	0%	0%	0%	0%	61%	1%	75%
Bus	0%	1%	0%	0%	0%	26%	71%
Recall	94%	25%	91%	92%	61%	26%	

Table 9: Confusion matrix with large spatial distortion for the pure classification problem.