Automatic update of road attributes by mining GPS tracks

Karl van Winden *
Delft University of Technology, The Netherlands
Filip Biljecki
Delft University of Technology, The Netherlands
Stefan van der Spek
Delft University of Technology, The Netherlands

ORCID
KvW: http://orcid.org/0000-0002-1041-6419
FB: http://orcid.org/0000-0002-6229-7749
SvdS: http://orcid.org/0000-0003-1258-0527

* Corresponding author at karlvw@hotmail.com

This is the peer reviewed version of the following article:

Van Winden K, Biljecki F, Van der Spek S (2016): Automatic update of road attributes by mining GPS tracks. *Transactions in GIS*, vol. 20(5), pp. 664–683.

which has been published in final form at

http://doi.org/10.1111/tgis.12186.

Abstract

Despite advancements in cartography, mapping is still a costly process which involves a substantial amount of manual work. This paper presents a method to automatically derive road attributes by analyzing and mining movement trajectories (e.g. GPS tracks). We have investigated the automatic extraction of eight road attributes: directionality, speed limit, number of lanes, access, average speed, congestion, importance, and geometric offset; and we have developed a supervised classification method (decision tree) to infer them. The extraction of most of these attributes has not been investigated previously.

We have implemented our method in a software prototype and we automatically update the OpenStreetMap (OSM) dataset of the Netherlands, increasing its level of completeness.

The validation of the classification shows variable levels of accuracy, e.g. whether a road is a one-way or a two-way road is classified with an accuracy of 99%, and the accuracy for the speed limit is 69%. When taking into account speed limits that are one *step* away (e.g. 60 km/h instead of the classified 50 km/h) the classification increases to 95%, which might be acceptable in some use-cases. We mitigate this with a hierarchical code list of attributes.

Keywords: data mining; GPS track; movement trajectory; OpenStreetMap; mapping

1 Introduction

With the increasing adoption of location-aware technologies such as smartphones, more and more movement data is becoming available benefiting a growing number of applications (Shoval et al., 2014). For instance, movement trajectories are being used in travel behavior research (Bohte, 2010; Van der Spek et al., 2009), health studies (Thierry et al., 2013), and traffic management (Ranjitkar et al., 2002). In GIS, movement data have been used for determining points of interest (Cao et al., 2010; Zheng et al., 2009), for detecting missing features (Heipke, 2010), and for the improvement of the geometry of features (Schroedl et al., 2004).

Research efforts coupling movement data and mapping are focused towards the geometry of features, and research investigating their semantic aspect (e.g. type of a road) is sparse. Our research, positioned at the intersection of data mining and GIS, attempts to bridge this gap by investigating if movement trajectories can be used to automatically derive and update attribute data in maps. Roads are one of the most prominent features in maps, and are directly associated with the movement of people, hence, we focus our research to the attributes of roads, such as their speed limit and directionality. Using movement trajectories for this purpose would potentially result in improving the efficiency of mapping production and the increase of the completeness of currently available datasets.

In this paper we investigate the automatic derivation of eight road characteristics: directionality (one-way or two-way road), speed limit, number of lanes, access for bicycles, average speed, con-

gestion, importance, and estimated error (offset) of the geometry of the feature. These attributes have been selected by exploring commonly available attributes in existing road datasets.

We have used a decision tree classifier after investigating patterns in trajectories, and validated it with a crowdsourced dataset recently acquired in the Netherlands with GPS. The algorithms that we have developed are not complex and can be reproduced easily, benefiting their implementation. They involve a number of thresholds which have been determined with a Monte Carlo simulation to find the optimal values of thresholds that minimize the errors. The dataset that we have used comprises a substantial number of GPS samples: it contains the movement of 800 people sampled every 5 seconds over 7 days in the Netherlands, hence it covers a wide travel behavior and geographic coverage. This is especially important for the three attributes (average speed, congestion, and importance) that reflect the actual usage (transportation flow) of the road. The size of this movement dataset is comparable to similar research campaigns in other countries, such as the "GeoLife GPS Trajectories" dataset from Microsoft Research (Zheng et al., 2009), which are publicly available.

To test our method with a real-world map dataset, we focus on OpenStreetMap (OSM) which is, at least in Western Europe, currently the most significant example of a system based on crowdsourced geodata (Girres and Touya, 2010), and it has been a topic of several research papers in GIS (e.g. Barron et al. (2014); Mooney and Corcoran (2014); Hochmair et al. (2015)). After deriving the attributes of roads in the Netherlands, we update the corresponding roads in the OSM dataset. The results show that a significant number of roads covered by the campaign are updated, increasing the OSM level of completeness. Further, some of the attributes are not present in OSM, hence we add them to OSM. OSM is not only used for the update, but also for validation, for which we use the existing data, which is covered well for the Dutch roads.

In Section 2 we present related work in this field, and our motivation for using OSM as the map dataset to be updated. Section 3 presents the methodology and related matter such as the used data, tools, and preprocessing. For seven attributes we develop a method for their automatic derivation. One attribute, the number of lanes, could not be automatically derived, and we present our efforts and point to the problem, supporting potential future work. The results and the validation of the method, along with the discussion of privacy concerns, are presented in Section 4. Section 5 concludes the paper and gives directions for future work.

2 Background

2.1 Related work

One of the similar efforts in this field of research is the work performed by Zhang et al. (2010), which investigates the integration of GPS tracks with road maps. The GPS tracks are used to improve the accuracy and detail of an out of copyright road map (e.g. OSM). Their research primarily improves the geometry of roads, but it also works on extracting attribute information such as directionality and turning restrictions. The research relies on the characteristics of the spread of the tracks from the center line of the road, which is modeled as a Gaussian distribution. In our

method we have used some of the insights, such as the importance of determining the center line of a road.

Chen and Krumm (2010) use a probabilistic method to derive the number of traffic lanes from GPS tracks. To find the lane structure, they "fit a Gaussian Mixture Model (GMM) to the intersections between the GPS traces and a sampling line perpendicular to the road's centerline". This is similar to the previously described research, and we attempt to reproduce their method (more on this in Section 3.3.4).

Li et al. (2014) use data mining techniques to extract the road class and road name from a combination of movement trajectories and geotagged data in social media. The road network is extracted using an incremental generation method, where road classes are obtained using Support Vector Machine (SVM). Their research is of our interest also because it provides insights into using SVM for developing classifiers for movement data.

Ekpenyong et al. (2009) present a research "towards an innovative solution to the problem of automated updating of road network databases". Their research identifies roads and assigns them into classes using a snap-drift neural network to analyze the GPS trajectories. A grouping accuracy of 71% is achieved. The research includes statistical indicators of trajectories (e.g. sinuosity and acceleration) that we consider in the design of our method.

The presented related work shows that the topic has not been researched to the full extent, and it is mostly limited to one or two attributes. In our work we consider a large number of attributes (eight), and we use the insights obtained from the review of related work to infer additional attributes.

2.2 OpenStreetMap

The OSM project represents the free editable map of the world. The positional accuracy of OSM dataset has been researched in several countries in Western Europe, for instance, France (Girres and Touya, 2010), United Kingdom (Haklay, 2010), Germany (Zielstra and Zipf, 2010), Ireland (Cipeluch et al., 2010), and the Netherlands (Bhattacharya, 2012). Researchers generally conclude that the completeness and positional accuracy of OSM data in Western Europe is comparable to government datasets. However, the quality of the attribute data is not consistent. It is our experience that the attributes suffer from low completeness, sometimes also lesser accuracy, providing a motivation for this research.

Attributes in OSM are described using tags. OSM has a freeform tagging system with key-value pairs, which allows the contributors to add an unlimited number of attributes to a feature (Haklay and Weber, 2008). However, certain tags and their values act as informal standards agreed upon by the OSM community.

Consequently, we have decided to use OpenStreetMap as the map dataset to test our developed method. OSM is free, hence we have the possibility to modify it, and the geometry is relatively accurate, which is important for our research since we aim to improve the semantics rather than the geometry. Finally, because of the flexible rationale regarding the attributes, they can be freely

extended with new keys and tags, which we do for a few attributes not previously available in OSM (e.g. average speed).

3 Methodology

This Section describes the developed method with our implementation, and insight into related topics such as the used data and tools.

The method consists of four steps: (1) acquisition of the movement data; (2) its preprocessing; (3) automatically deriving the attributes of roads; and (4) updating the map dataset with the derived attributes (the latest step is described in the next Section 4 with the obtained results). The algorithms have been developed by analysing a training dataset (Section 3.3.1). The steps are shown in Figure 1, and are explained throughout this Section.

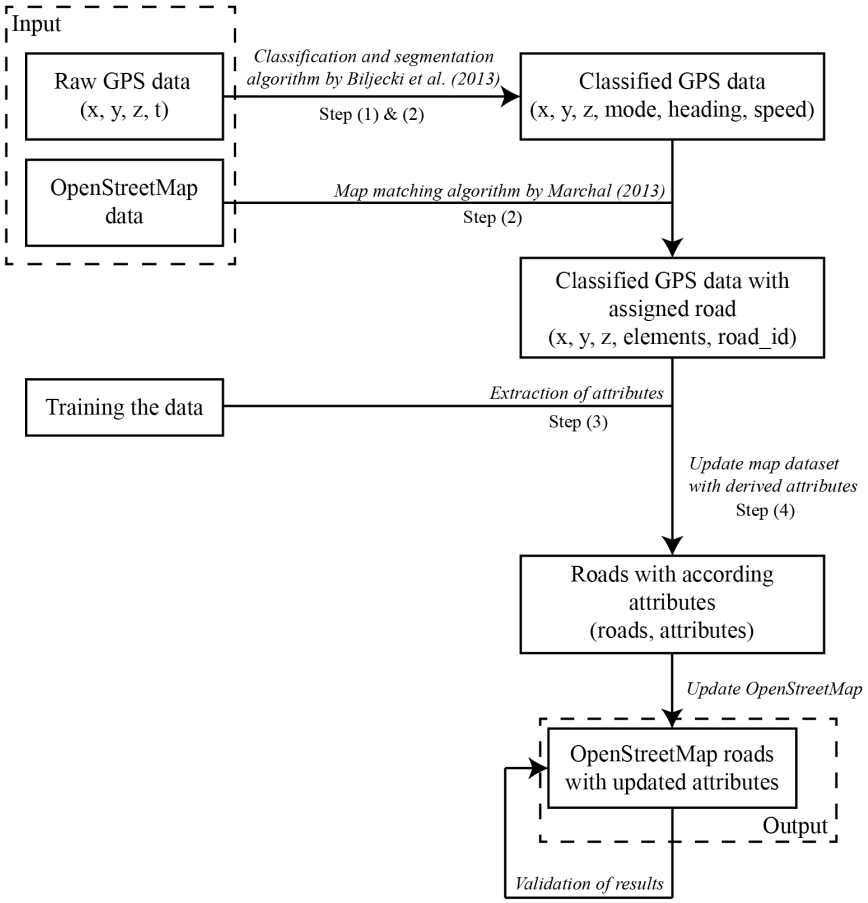

Figure 1: Diagram explaining the methodology of the research.

The implemented software prototype uses a spatially enabled PostgreSQL database that contains the map data (OSM) and the movement data (GPS tracks). All attribute extraction algorithms are implemented using Python.

3.1 Acquisition of movement data

The movement dataset (i.e. GPS tracks in the form of a set of timestamped positions x, y, z, t) was acquired in early 2013 for an urban analysis project at the Section Urban and Regional Development of the Delft University of Technology in the Netherlands (Van de Coevering et al., 2015). The dataset contains timestamped positions by more than 800 people inhabiting the cities of Amersfoort, Veenendaal and Zeewolde (central Netherlands) during a 7 day campaign in which respondents carried a GPS device. The respondents of the campaign have volunteered to participate in the survey after they have been approached randomly. The distribution and density of the samples across the Netherlands is shown in Figure 2.

Such dataset is comparable to similar campaigns, which are also available for free public use across the world (e.g. see the work of Zheng et al. (2009)). An alternative is to use the OSM GPS data, uploaded by users worldwide.

3.2 Preprocessing the data

The preprocessing of the movement data consists of four stages.

First, because the algorithms which we have developed require additional data such as speed and heading at each sample in the track, each of the point is enriched with a series of calculated parameters.

The method takes advantage of mining trajectories made by car and bicycle. Since the data contains tracks made by a variety of transportation modes, knowledge of the used mode is important for filtering out tracks made by non-relevant modes.

The classification of the movement trajectories regarding the used transportation mode has been done with the method developed by Biljecki et al. (2013). The method was selected because of its accuracy and available implementation. After the classification, each GPS point is enriched with the information of the used transportation mode in that moment.

The third stage is focused on the map matching of the GPS points, i.e. assigning each point to the road. This part is essential for deriving the attributes for the roads because it is required to know to which road each GPS point *belongs*. The algorithm that is used for this research is based on the topological algorithm developed by Marchal et al. (2005), and it was realized in a cloud based web service TrackMatching*. The motivation for using this approach is its accuracy and the free availability of the implementation.

*https://mapmatching.3scale.net

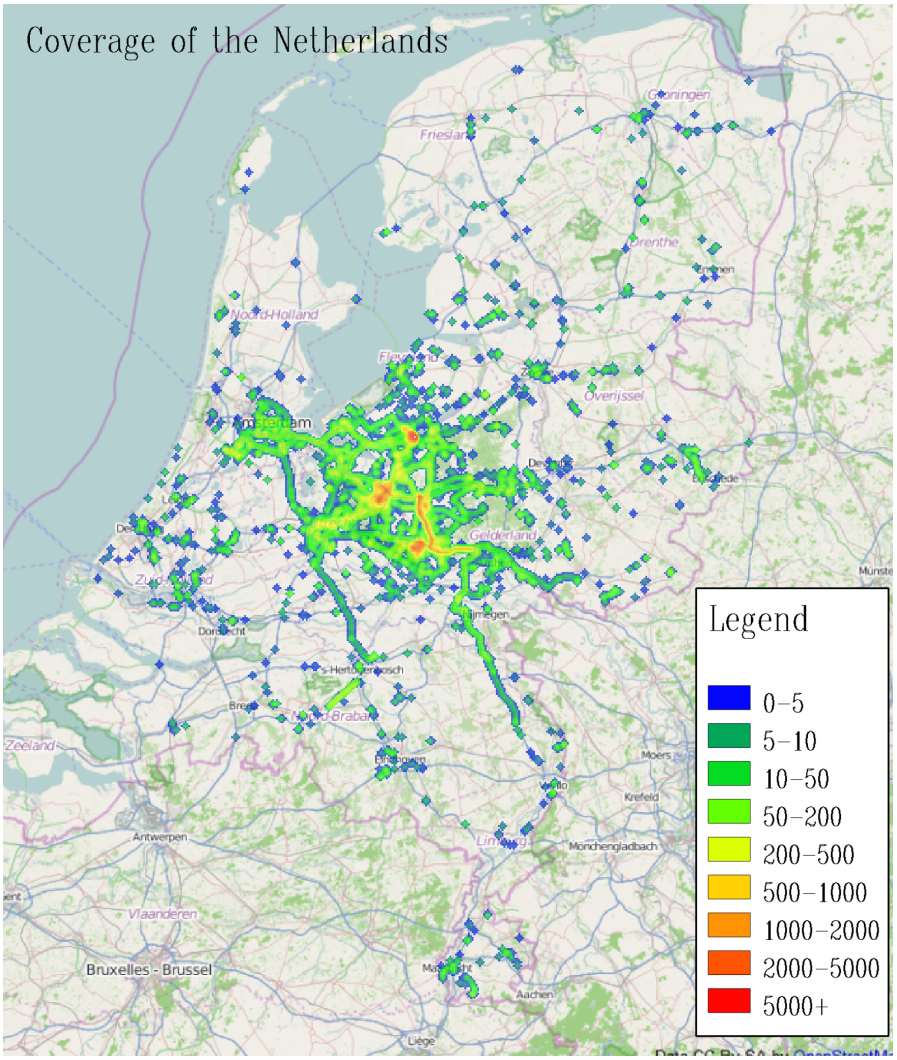

Figure 2: Density map of the movement dataset showing the distribution of the GPS samples in the Netherlands. Since the country is not fully covered with a uniform number of samples (explained in Section 3.3), the attributes have not been derived for the whole country. The colors and the legend indicate the number of points per square kilometer.

In the fourth stage biased movement data which might negatively influence the quality of the results are filtered out. We did this by imposing several thresholds to mitigate frequent problems, such as undesired travel behavior and GPS noise. For instance, in several occasions during the campaign people forgot to take the GPS device with them or have stayed at home with the device on, resulting in many points at the same location. These points are all matched to the same road, thus it appears that the road is used a lot, and that the traffic on it is congested. Therefore some measures have been taken to filter out these points, such as taking into account the duration of the stationary period and the speed of each point, e.g. all points that have a speed lower than 5 km/h

are removed from the dataset. Another example are outliers, which are detected by observing unrealistic accelerations.

3.3 Deriving Attributes

A decision tree algorithm has been developed for deriving each attribute. A list of possible classes per attribute has been determined by following the standard OSM practices. However, in order to fully relate our method to OSM, we extend OSM in two ways.

First, multiple code lists have been implemented in a hierarchical structure, which we call a hierarchical code list. The hierarchy describes the level of detail and the granularity of the value of the attribute, and it is introduced for accuracy reasons. This is in line with the general notion of the level of detail in GIS (Biljecki et al., 2014). Due to the complexity of the derivation of some attributes, the hierarchies can provide different perspectives on the classification. For instance, the classification of the speed limits for fine classes proved to be complex and may lead to inaccurate results. Hence, we have decided to group classes resulting in a coarser classification, which might be acceptable for some use-cases.

Second, in practice OSM does not contain four out of eight attributes that we investigate (average speed, congestion, importance, and geometric offset), hence we add new keys to OSM that are currently not present.

The research that has been done for deriving each attribute is described in the following sections, with an algorithm, which is described either in pseudo-code or narratively. Further, for each attribute, we have determined the minimum number of points that should be used in order to get satisfactory and unbiased results of the classification. This was determined by running the algorithms in iterations with different sizes of randomly selected subsets, and finding the degree at which the increase of the number of samples has no further influence on the accuracy of the classification. The algorithms, except the one for the congestion (Section 3.3.7), are independent of each other because of the specific challenges that are unique to each attribute (i.e. each procedure is different and it does not require the knowledge of other attributes).

3.3.1 General methodology and training dataset

For designing the classification tree for each attribute, we have followed a general approach that we elaborate here.

First we have established a training dataset to discover the potentially predictive relationships between the movement trajectories and the values intended to be acquired. The training dataset consists of a number of road segments for which the ground truth of the attributes is known (from OSM and confirmed with official datasets), and the movement trajectories that are matched to them (subset of the dataset described in Section 3.1). The selection of the training dataset will be explained after the general methodology.

For each of the road segments we have computed a number of statistical indicators from the matched movement trajectories that *might* indicate the value of the attribute. For instance, for the attribute of the speed limit, the mean speed, the median speed, and its distribution, have been computed. The selection of these indicators is in line with the methods of related work (Section 2) and in the work on the classification of movement trajectories for other semantic properties such as the transportation mode and purpose (Stopher et al., 2008; Bohte and Maat, 2009; Biljecki et al., 2013).

The statistical indicators have been analyzed, and the relation to the values has been established through a classification tree. The thresholds that have been used are determined with a Monte Carlo simulation on a set of thresholds from which the optimal value yielding the smallest error has been found.

In this method, the training dataset plays an important role. The training dataset has been selected to be representative with respect the different types of roads, travel behavior, and geographic area, and as such has to contain the attributes that are investigated (the values in OSM have been checked with other sources such as government datasets). Following this rationale we have chosen the road segments that contain most of the samples in our dataset, and where different categories of roads are equally represented and in different geographic areas. This covers a wide travel behavior and different types of roads.

The road segments are structured according to the data model of OpenStreetMap, where each road may be represented by multiple line segments. In our training dataset, the road segments are of varying distances from 0.5 to 5 km, and they contain from 6 to 19 thousand samples.

The number of 10 road segments have proven to be of a sufficient size. This has been tested with repeated experiments involving different permutations and sizes of the training dataset. We have computed several statistical indicators, and the results did not significantly deviate by increasing the number of road segments.

3.3.2 Directionality (one or two way road)

The knowledge of whether a road facilitates driving in one direction or in two directions is critical for routing and other purposes. This attribute is realized in OSM with the "oneway" tag, with values of "yes" and "no".

The algorithm first selects the GPS points that belong to a particular road thanks to the map matching performed in the preprocessing step (Section 3.2). Then using the heading of each road (in degrees), the points are grouped into three bins: "similar", "opposite", and "outliers". The "similar" points comprise the points whose heading is approximately the same as the corresponding line segment that represents the road. The reverse applies to the "opposite" points. There is no "equal" class because the heading of two consecutive GPS points always has a deviation from the line segment on the road.

The threshold to determine if a point has a comparable heading is 20 degrees. All points that are outside this threshold are outliers and are not taken into consideration. As with other thresholds

used in this work, this value has been determined with repeated experiments (as a Monte Carlo simulation) in which the optimal threshold has been determined. More details can be found in Van Winden (2014).

The total number of "similar" (or "opposite") points is divided by the sum of "similar" and "opposite", and if this ratio is bigger than a certain threshold the road is considered a one way road. Else, the road is a two way road. The algorithm is tested by using different thresholds and the threshold with the smallest error on the training data was selected. In our experiments the ratio of 0.9 proven to be the best value. This means that if a specific relative direction is represented less than 10% the road is considered one way accounting for undetected outliers.

The minimum number of points required for this classifier is 10 per road, which was determined with experiments when taking into account the different number of samples.

3.3.3 Speed Limit

In routing, also an important attribute of roads is its speed limit, which in OSM is realized through the "maxspeed" tag. The assumption here is that by deriving the movement pace from the trajectories it is possible to deduce the speed limit of a road. Deriving the speed limit depends on different factors. For instance, the varying behavior of drivers has a considerable influence. Further, speed limits can differ over time, and congestions may cause significant bias. Such aspects influence the input data and make it difficult to develop a reliable reasoning for automatically and precisely deriving speed limits. Here we briefly describe the steps that we have taken to create an algorithm that derives the speed limit from movement data.

The "maxspeed" attribute can be divided into multiple levels in the hierarchy (Figure 3). The speed limits that are used here are the speed limits in the Netherlands: 30, 50, 60, 70, 80, 100, 120, and 130 km/h. Level L0 groups these speed limits into an ordinal scale: low, medium and high.

By analyzing the trajectories, we have realized that there is a clear difference in driving behavior between two roads with the same speed limit, e.g. the average speed of samples on one road is different from another road with the same speed limit. For lower speed limits, the deviation of the behavior is less noticeable.

In the movement data there are also many relatively low speeds which could influence the results of the extraction of the speed limit. These relatively low speeds can be considered as outliers and are caused by deceleration (e.g. for stopping at traffic lights), congestions, and other forms of traffic disruptions. Therefore it is important to filter out lower speeds which do not represent the uninterrupted flow of the traffic. Using acceleration proved not to be beneficial. Hence, the velocity change rate (VCR), a concept different from acceleration, has been used. It is a statistical indicator that takes into account the magnitude of the velocity, i.e. a change of 20 km/h from 100 km/h is different than the same change from a speed of 30 km/h. VCR is calculated by dividing the change in velocity by the velocity of the first point (Zheng et al., 2010):

$$v_{\Delta} = \frac{v_i - v_{i-1}}{v_i} \tag{1}$$

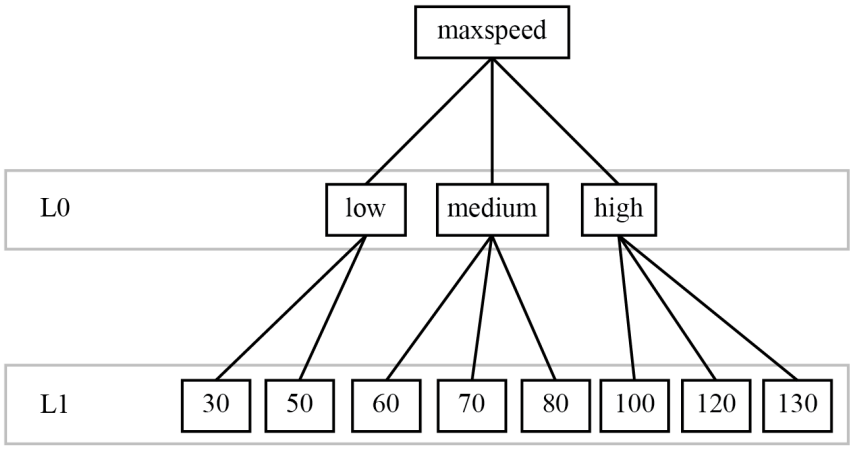

Figure 3: Different levels in the hierarchical code list for attribute "maxspeed". The units are in km/h. This hierarchy may slightly differ between different countries where the speed limits and units may be different.

Our research has shown that 0.15 was the threshold which gave the best results. Therefore only points with a VCR between -0.15 and 0.15 are taken into account when deriving the speed limit.

After filtering the non-relevant points, we have realized that the mean speed of the points on a road is proved to be a valuable indicator for the speed limit on the road. A step function of speed limits was created taking the mean of the speeds as its single argument. Investigations based on the training data have shown that the classification according to Table 1 yields the best results.

Mean (input) [km/h]	**Speed Limit (output) [km/h]**
< 40	30
> 40 and < 55	50
> 55 and < 65	60
> 65 and < 75	70
> 75 and < 85	80
> 85 and < 110	100
> 110 and < 125	120
> 125	130

Table 1: Classification function of the speed limit according to the mean speed. This hierarchy is country-dependent depending on the legislation and on the units.

However, because of varying behavior in the different speed limit categories, for the higher speeds it proved more difficult to classify the correct values using the standard value of mean. To solve this problem, we keep the presented classification function, but when the mean speed is higher than 85 km/h we do an additional filtering where only a percentile of the highest speeds is taken into account. Experiments did show that the 20th percentile was the best value to use for deriving

the speed limits above 85 km/h.

Afterwards, our investigation has shown that only using the mean speed for speed limits lower than 85 km/h and only using the mean of the percentage highest speeds for speed limits higher than 85 km/h is not sufficient. The system inclined to assign lower speed limits to the roads. Therefore to improve the performance of the algorithm, a mix of the two means is created. For speed limits up to and including 80 km/h, the average of the mean speed (v) and the mean of the percentage of highest speeds (v_{high}) is used: $\frac{v+v_{high}}{2}$. For speed limits higher than 80 km/h, the mean of the percentage of highest speeds is weighted twice than the mean speed: $\frac{v+2v_{high}}{3}$. This is because the higher speeds are more representative than the lower speeds for high speed limits. These values are tested on the training dataset (described in Section 3.3.1) to determine the best weight for this classification.

The final algorithm, which requires at least 50 samples on a road segment, is shown in Algorithm 1.

Algorithm 1 Attribute Extraction of "maxspeed".

```
function MAXSPEED(roadid)
    Select all speeds for roadid
    if r ≥ 50 then                                  ▷ r is the number of results
        Derive mean from speeds
        SPEED LIMIT(mean,speeds)
        Update "maxspeed"

function SPEED LIMIT(speeds)
    if ((v + h)/2) ≤ 40 then                                  ▷ v = normal mean
        maxspeed = 30                                  ▷ h = high speeds mean
    else if ((v + h)/2) > 40 and ((v + h)/2) ≤ 55 then
        maxspeed = 50
        etc...
    else if ((v + h)/2) > 85 then
        if ((v + 2h)/3) > 85 and ((v + 2h)/3) ≤ 110 then
            maxspeed = 100
            etc...
    return maxspeed
```

Inspired by related work (Section 2), besides this algorithm, some pattern recognition classifiers such as SVM were also trained on the data but were proven not to be more accurate than the current (simpler) method.

3.3.4 Number of usable lanes on the road

This Section discusses the possibility of the detection of the number of usable traffic lanes of a road. The difficulty with GPS data for lane extraction is its inaccuracy, noise and the position of

the GPS device with respect to the lanes.

Chen and Krumm (2010) use a histogram with the assumption that the distances of the samples to the centerline of the lane are normally distributed, and develop an algorithm that extracts the number of lanes. We could not manage to reproduce their method on our data following their publication, and we have worked towards alternative solutions, which did not succeed either, as experiments on our data suggest that the lanes are indistinguishable. However, we describe our efforts in order to support possible future attempts.

First it is important to determine on which side of the road a point lies. This can be determined using a cross product, and it is done for each point in the dataset. Afterwards, the distance to the center line of the road has been computed.

For an impression of the calculated data, a histogram of the distances from each point to the center line of one direction of a road is shown in Figure 4. Even as a human, it seems impossible to detect the number of lanes out of this histogram.

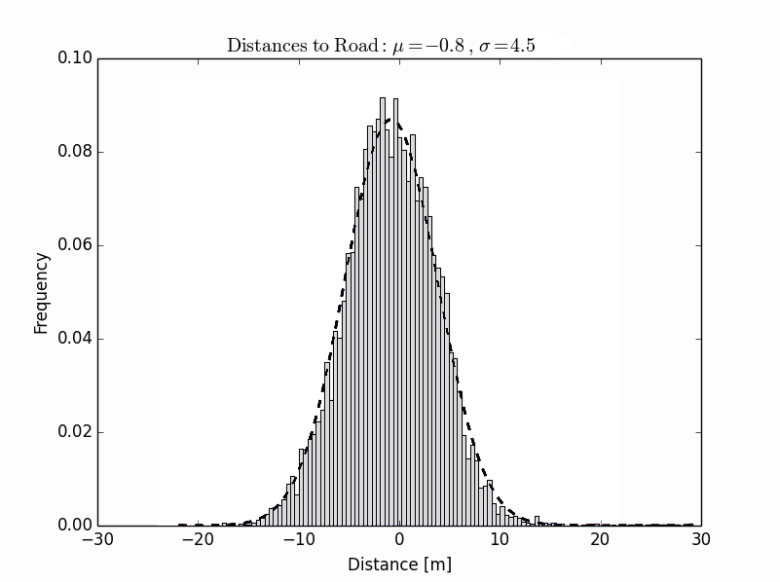

Figure 4: Histogram showing the distances from GPS points to the center line of the road including the fitted normal distribution.

An alternative approach to infer the number of lanes would be from the width of a road, however, the standard lane width in the Netherlands substantially varies (between 2.5 meters and 4.5 meters depending on the road type (WegenWiki, 2014)). Even when using these lane widths and the standard deviation as an indicator for the spread of the points, one might expect that the number of lanes can be derived. However, calculating the mean and standard deviations did not contribute towards a successful classification. Further, the shape of the distribution does not reveal much, as it is a frequent case that the distribution of distances of samples on roads with two or more lanes are *blended* in a single *peak*.

We have computed histograms and values for each road in the training dataset, and we have not seen considerable deviations in the values and in the distribution of the distances when analyzing roads with a different number of lanes.

According to Chen and Krumm (2010), Gaussian Mixture Models (GMM) can detect lanes from GPS data. Reynolds (2009) defines GMM as "a parametric probability density function represented as a weighted sum of Gaussian component densities". Implementing GMM and testing it on the training dataset, did not result in a successful lane detection either.

Hence, despite previous successful research, we conclude that when dealing with real-world data the number of lanes cannot be determined from movement trajectories. We are of an impression that such methods work only with data obtained in a controlled environment, for instance, one where highly accurate devices are mounted at the same position in the vehicles.

This inability is most likely caused by the combination of multiple factors that influence the movement data. Not only the GPS noise considerably affects the performance of the classification, but also the placement of the devices in a vehicle can cause deviations in this case. We believe that with the improvement of the accuracy of positioning sensors in future, this problem might become easier to solve.

3.3.5 Access

The term access in the frame of OSM has a two-fold meaning, both related to the transportation mode. On one hand, it can generally denote a traffic restriction for a certain road, i.e. that no non-walking transportation mode is allowed on a particular road (e.g. it is strictly a pedestrian zone). On the other hand, not all types of vehicles are allowed on every road, therefore it is used to denote the different categories of vehicles (e.g. truck, car, and bicycle) that are allowed. This can be useful in navigational systems, which can then determine appropriate routes for each mode (e.g. different for a car and a bicycle; see the related research of Hochmair (2005) and Kim et al. (2009) on deriving the appropriate route selection based on different factors).

From the implementation point of view, the first type of access cannot be taken into account because in reality it is not possible to obtain a movement dataset that is so dense and thorough where road segments without any recorded movement could be easily detected and assigned as restricted. For instance, if no pedestrian GPS samples are recorded in a restricted (pedestrian-only) street, it would be ambiguous if such street is a pedestrian zone or a private road that is off limits. For these reasons we focus on the latter, where we extract the restrictions related to a class of a non-walking transportation mode. We have developed an algorithm that analyzes the position of a track made by a bicycle relative to the nearest cycleway and road, and it determines if bicycles are allowed access to a road.

The development of the algorithm is hampered by the practical fact that in practice in OSM cycle paths are not mapped as much as roads, and that the map matching implementation that we use does not support matching to cycle paths. Hence, we adapt the algorithm to mitigate these shortcomings.

The method for the classification is given narratively, and it requires at least 10 samples.

First for each point a motorized road is matched and a nearest cycleway is assigned, provided that there is a cycleway in the vicinity. This algorithm takes into account a couple of factors. At first, a GPS point made while cycling has to comply with the same thresholds as the GPS points made by car, i.e. a speed higher than 5 km/h, the distance to the road should be less than 30 meters and the heading of the point should be similar or opposite to the drawing direction of the road. Secondly, the distance to the road should be smaller than the distance to the cycleway. Finally, the difference between the distance to the road and the distance to the cycleway should be bigger than 10 meters. The latter should be incorporated into the system due to points that are slightly closer to the road than the cycleway, but actually are on the cycleway, which is caused by GPS noise.

3.3.6 Average Speed

The average speed of the traffic on a road is an attribute not found in OSM, and it might be estimated when dealing with a large sample of movement trajectories. It an attribute that is relatively easy to calculate, and is frequently used for navigation purposes, for instance in routing when querying for optimal routes.

We introduce this attribute to OSM as "averagespeed". While a single value per each road might be calculated, we develop a finer code list. The level L0 of "averagespeed" is the average speed of a road in both directions. Level L1 is the average speed for each direction, in case of a two way road. Finally, level L2 provides the average speed per direction per hour to take into account deviations of speed during the day for more detailed information. The levels in the hierarchy are depicted in Figure 5.

The average speed in level L0 is calculated by selecting all the speeds of GPS points of a certain road and dividing them by the total number of GPS points. For level L1, only points with the relative directionality "similar" or "opposite" are selected, and for level L2 all the points with the relative directionality and a specific hour are selected. We have realized that for all the levels this algorithm gives results with only 10 points (increasing the number of samples did not increase in the accuracy of the prediction), and when running the classification with an increased number of the samples, there is an insignificant deviation (i.e. <1%) from the values obtained with 10 points.

3.3.7 Hours in which congestion occurs

The time of day during which a road is usually congested might also be useful for navigation purposes, to avoid certain roads at certain hours. Taylor et al. (2000) describes congestion as a phenomenon of increased disruption of traffic movement on an element of the transport system, observed in terms of delays and queuing.

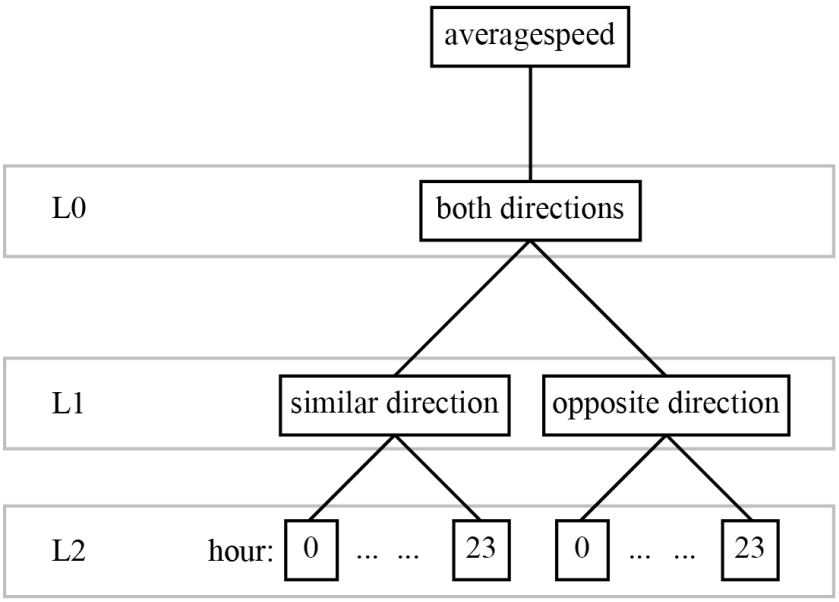

Figure 5: Different levels in the hierarchical code list for attribute "averagespeed".

We introduce this attribute to OSM through the tag "congestion", which indicates the usual congestion for a specific hour of day. The level L0 represents the difference between congestion during the week and the congestion during the weekend. Congestion tends to happen more often during weekdays, because of the working hours of inhabitants. Therefore level L1 separates the congestion per day of the week (168 classes). The levels in the hierarchical code list are depicted in Figure 6.

There are two possible ways to derive the hours in which congestion occurs. One method is to compare the average speed of a specific hour to the average speed of the road and the other method is to use the VCR to detect a significantly higher amount of stopping or slowing down of the car compared to the uninterrupted behavior.

In this research, we apply the first approach because it achieves the goals of congestion more directly. Moreover, stopping at traffic lights and differences in speed limits on roads could cause the VCR to fluctuate making it not reliable for detecting congestions. Hence, we define disruptions as considerable deviations from the average speed of a road.

First of all, it should be noted that the congestions are calculated per driving direction. This is important because it is possible that in one direction there is no congestion, but in the opposite direction there is a lot of congestion during a certain moment in time. This means that for the opposite direction a different route might be more beneficial, but not for the other direction. Next, a percentage of the average speed is determined as the indicator for congested traffic. In this case, 10% seemed as an optimal percentage to define as a significant difference between the average speed of the road and the average speed per hour. We have determined this by analyzing the pace

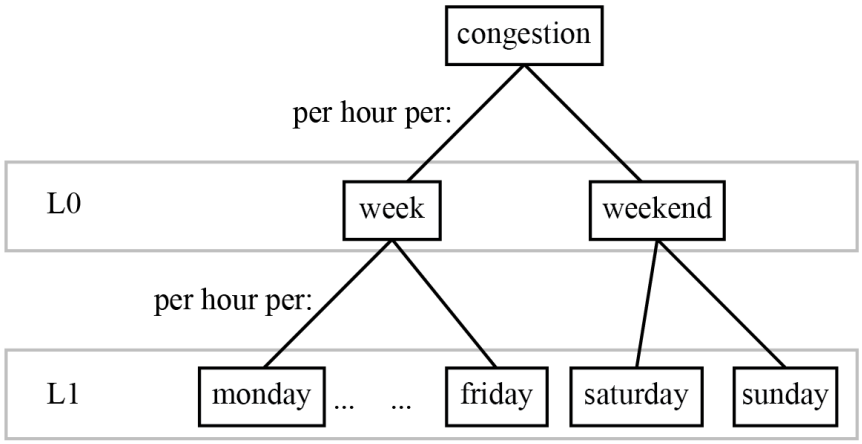

Figure 6: Different levels in the hierarchical code list for attribute “congestion”.

of the trajectories. The average speed per hour is then subtracted from the average speed and compared with the 10% of the average speed of the road. If the difference between the average speed and the average speed per hour is exceeding this threshold, that hour will be considered congested.

For a proper derivation of the values, a minimum number of 10 points per road direction per hour is recommended. As with other values, this was determined with a Monte Carlo simulation with a different number of samples.

3.3.8 Importance of a road

Generalisation of maps, navigation systems, and other applications require the knowledge of the classes of roads. While the importance of roads can be based on their type, this does not take into account their actual usage and demand. The distribution of acquired tracks for each road can give an indication of the usage of roads. Therefore, importance in this research can be described as the relative usage of roads compared to the total usage within the movement dataset. Depending on the coverage of the movement trajectories, this may refer to cities, regions, and countries (see Figure 2 for the kernel density estimation in our case).

Therefore we introduce the “importance” attribute, which indicates the usage of the road on an ordinal scale. There are multiple ways to derive the usage of a road. The basic idea is to derive the ratio of the usage of the road compared to the total usage of all roads. The following units can be used to derive the importance: number of GPS points on a road, time spent on a road by GPS points, and number of passes on a road.

The first unit is the simplest: the more people drive on a road the more GPS points that road contains. However, in the case of regularly sampled data, this unit also has a disadvantage: the

varying speed. The speed influences the number of GPS points that are assigned to a certain road and it biases the result.

By taking into account the time spent on the road, the results will also be biased in the same manner. For example, consider motorways which are often congested. Such roads will then have a higher amount of time spent on them, while they may not actually be used more often. Hence the time unit also does not represent the usage of roads well.

The number of passes could give a better, less biased result compared to the previous two units. A pass is counted when a vehicle travels over a certain road, and it is equal regardless of the time spent and the speed. Therefore we have used it as a measure for the importance.

After the number of passes on all roads are calculated, a classification is performed using a step function where a range of the number of points is assigned to a single class. This classification depends on the data and the amount of passes that are made. Because of a non-uniform distribution of points, we have decided not to assign directly the number of passes as the value, but rather an ordinal scale of five classes that are determined from an evenly sized bins of points—the amount of the highest number of passes in the computations has been divided by five to represent the size of each class (e.g. 1001-2000 and 2001-3000).

3.3.9 Offset of the Road Geometry

The error of the geometry of the road network in OSM is usually assessed by a comparison to official datasets. However, here we introduce the possibility of using movement trajectories for this purpose. This attribute can therefore be considered as a quality indicator.

A multitude of research projects is available where road networks are extracted or improved from GPS points. For example, the research by Zhang et al. (2010) improves existing road data from GPS tracks. Similarly, Bruntrup et al. (2005) uses GPS tracks to generate maps and infer the road geometry. However, they do not to use the GPS points as a quality indicator. The advantage of GPS tracks is that, while they are not too accurate, if the dataset is big enough eventually a line derived from a collection of tracks will follow the geometry of the road.

Here we consider a different perspective and try to determine the positional accuracy of the road geometry, store it as an attribute, and make it possible for OSM contributors to be alarmed about excessive offsets.

Again, the relative distance is used to calculate this attribute. The mean of all relative distances of a road is the average offset of the GPS data compared to the center line of that road. Figure 7 shows an example for one road. The dashed line can be considered as the approximate centerline of the trajectories, and the origin value (at 0) is the OSM centerline. Their discrepancy may be a useful indicator of the accuracy of the geometry in OSM.

However, in some cases, this value cannot be directly used as a quality indicator due to the fact that in the case of multiple lanes the distribution of the GPS points over these lanes influences the result. For example, if for a two lane road 90% of all the people on that road drive on the

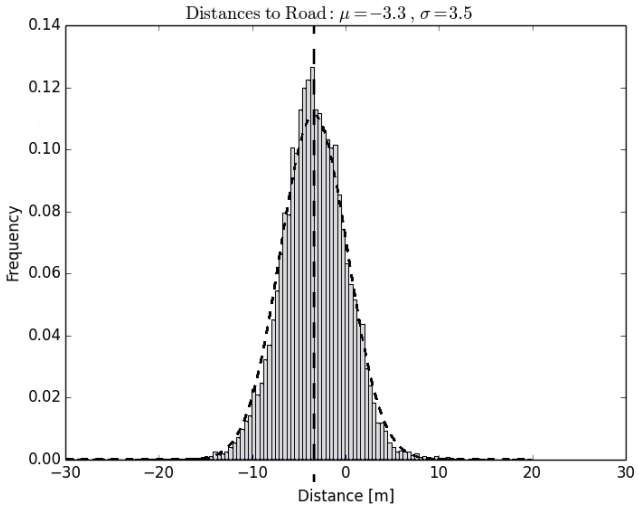

Figure 7: Histogram of a road showing the discrepancy between the OSM geometry (at 0 meters) and the mean of the distances of the samples (at -3.3 meters—dashed line).

right side and 10% of the people drive on the left side, the mean of the relative distances will be shifted and will not represent the center line of that road. Further, the distribution of traffic in two-way roads may differ, which could also have a substantial influence. Since it was not possible to derive the number of lanes, it is not possible to adjust the distribution of points with respect to the number of lanes. However, since such significantly uneven usage of the lanes are common only in highways (e.g. ones having three or more lanes), this attribute still serves its purpose for most cases.

4 Results, validation and privacy

4.1 Updating the attribute values

After the road attributes have been derived, the map dataset has to be updated. In our case, because of the experimental nature of this research, the derived attributes were updated in a local copy of the OSM database.

In Table 2 we present the new level of completeness of the attribute data of the roads in the area covered by our movement data, i.e. roads that have at least one sample matched to it. Depending on the attribute, the completeness varies from 4% to 100%. For instance, now 60% of the roads contain the value for the directionality. This represents the increase of 39% comparing to the presently available data.

It is important to state that the Netherlands is one of the most completely mapped countries in OSM (a fact that is also visible in Table 2). Hence, when using a movement dataset of a comparable

size, the increase in completeness for other countries would be more substantiated. Further, the high level of completeness in the Netherlands is beneficial for this research because we have used the existing data for validation (Section 4.2).

Two attributes: "access" and "congestion" have been derived for only a fraction of the roads because the used dataset did not have enough samples for their extraction on a larger scale.

Attribute	**Informed roads**		**Relative increase** [%]
	before the extraction [%]	**after the extraction** [%]	
"oneway" (§3.3.2)	43.1	60.1	39.4
"maxspeed" (§3.3.3)	50.8	52.5	3.3
"access" (bicycle) (§3.3.5)	0.0	4.0	–
"averagespeed" (§3.3.6)	–	37.3	–
"congestion" (§3.3.7)	–	5.5	–
"importance" (§3.3.8)	–	100.0	–
"geometryerror" (§3.3.9)	–	34.1	–

Table 2: Overview on the magnitude of the informed attributes in OSM before and after performing our classification.

4.2 Validation

Here we provide an overview on the accuracy of the classification of the attributes, an analysis of the errors, and discusses the sources of the errors. Since there is a number of external factors that influence the performance of the classification, the validation focuses both on the total error of the classification, and the identification and elimination of the erroneous external factors.

The validation was performed on the subset of 100 roads for which the ground truth is known. The ground truth has been determined from OSM, and it has been further checked with other sources such as government datasets. A number of attributes is novel (e.g. average speed and congestion), and for them the validation against ground truth is not possible.

An overview on the classification accuracies is given in Table 3. The Table contains both the total error of the classification, and the errors of the classification system after the eliminated external factors, which is explained later.

The error for determining the directionality of the road is 0.6% for both levels in the hierarchical code list. The error for the speed limit is 10.4% for the level L0 and 34.0% for the level L1, showing the difficulty of the classification of the speed limit in the finer code list.

However, if we take into account speeds that are one speed class away (e.g. classification of 50 km/h instead of the actual 60 km/h) the error is decreased to 5.0%. This may be acceptable for some use-cases, such as the rough estimate of travel time between two points. The access attribute has a total classification error of 26.0%.

Attribute	Level	Total accuracy [%]	Refined accuracy [%]
"oneway"		99.4	99.5
"maxspeed"	L0	89.6	89.6
	L1	66.0	69.2
	± one class	95.0	96.0
"access"	L0	74.0	90.2

Table 3: Overview of the accuracy of the developed classification system for attributes for which the ground truth is known.

The resulting errors are caused by different factors, which may be internal (algorithm-inherent) or external (data-inherent). Internal errors are the errors which are caused by the imperfection of the extraction algorithm developed in this research. External errors are errors which are present in the input data and cannot be easily detected. There are numerous external factors which can cause these errors. For instance, (1) wrongful classification of the transportation mode of the movement trajectory; (2) errors of the map matching, i.e. GPS points that are matched to another road; (3) data used as ground truth is not correct. This applies also to *temporary errors*, for instance, a road that has been closed due to constructions during the campaign, but it is available in OSM.

The errors caused by external factors have been filtered out with manual intervention (a sample of erroneous data has been manually inspected), and the refined results showing only internal errors of the algorithms are shown in Table 3. There is an improvement, especially for the "access" attribute.

4.3 Privacy

If the used movement dataset is sufficiently large (as it is in our case) there are no privacy issues concerning the movement and identification of people. First the data that is used has been anonymized and this research deals with investigating the utility of the GPS data, which means that no direct link between the person that carried the device and the data that is saved. Second, and more importantly, the algorithms derive single and unrelated values (e.g. speed limit) from which individual tracks, and further, their identity, cannot be reconstructed.

5 Conclusions and future work

In this article we have investigated to what extent it is possible to automatically derive road data attributes from movement trajectories. We have thoroughly researched eight attributes, which are common in road datasets. Four of these attributes are practically new to OSM. The algorithms that we have developed are straightforward, and can be easily reproduced. Because each of the attributes is specific, each requires a separate algorithm. We believe that it would not be possible to design a unified single approach to derive all atributes.

Further, we have shown that it is straightforward to extract some of the attributes (e.g. directionality), and for some it is difficult or impossible (e.g. number of lanes).

Our contribution is that (1) our comprehensive research involves a large number of attribute classes, most of them not researched previously; (2) we extend OSM by developing a hierarchical code list and introduce attributes that do not exist in OSM; (3) the presented methodology is fully automatic and it is implemented in a software prototype that derives the attributes starting from the import of the movement data up to and including the updating of the maps, without any human intervention; (4) we obtain satisfying results for some attributes (e.g. directionality and speed limit), and for the number of lanes prove that it is not possible to derive them from movement trajectories (at least not in the Netherlands); and (5) our method considerably elevates the level of completeness of the attributes in OSM, with some attributes being enriched in more than half of the roads covered by the trajectories. When updated, the attributes can contribute to a number of applications, such as navigation. Finally, OSM contributors can benefit from the new "geometryerror" attribute to get alerted of geometry that potentially has a significant offset from the real-world counterpart.

This research is not applicable to only the OSM map data and GPS data, but also for other map data and movement trajectories acquired with other technologies (e.g. GLONASS and Galileo).

While the method was tested in the Netherlands, it has a worldwide applicability after a certain degree of adaptation and adjustments of thresholds to conform to the travel behavior in other geographic areas. This depends on how similar the road network properties are in another geographic area, somewhere it would require minimum adjustments, and somewhere an additional training and a combination with other methods.

The classification system does not provide results with a 100% accuracy, but we are of the opinion that the accuracy cannot be further considerably improved, while retaining the same level of simplicity that our method offers. Potential improvements would require much more complex methods, and we have also concluded that some of the more advanced methods (such as SVM) do not bring any benefit.

For future work we plan to add a certainty measure to each classification, to extend the classification to other attributes which may be more challenging (e.g. type of the road and material of road surface), and to extend the hierarchical code list for some existing attributes, especially for the speed limit that can be different between day and night. Further, we plan to work on the validation of the attributes which were hard to validate in the scope of this research. This applies foremost to the attribute of the congestion, which can be validated with the use of Traffic Message Channel (TMC) data.

Acknowledgments

The helpful comments of the anonymous reviewers are gratefully acknowledged. We thank Paul van de Coevering for sharing the GPS dataset, and Fabrice Marchal for the access to the Track-Matching service.

This research is supported by the Dutch Technology Foundation STW, which is part of the Netherlands Organisation for Scientific Research (NWO), and which is partly funded by the Ministry of Economic Affairs (project code: 11300).

References

Barron, C., Neis, P., and Zipf, A. 2014. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS*, 18(6):877–895.

Bhattacharya, P. 2012. Quality assessment and object matching of OpenStreetMap in combination with the Dutch topographic map TOP10NL. Master's thesis, Delft University of Technology, Delft, the Netherlands.

Biljecki, F., Ledoux, H., Stoter, J., and Zhao, J. 2014. Formalisation of the level of detail in 3D city modelling. *Computers, Environment and Urban Systems*, 48:1–15.

Biljecki, F., Ledoux, H., and van Oosterom, P. 2013. Transportation mode-based segmentation and classification of movement trajectories. *International Journal of Geographical Information Science*, 27(2):385–407.

Bohte, W. 2010. *Residential self-selection and travel. The relationship between travel-related attitudes, built environment characteristics and travel behaviour*. PhD thesis, Delft University of Technology, Delft, the Netherlands.

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. *Transportation Research Part C: Emerging Technologies*, 17(3):285–297.

Bruntrup, R., Edelkamp, S., Jabbar, S., and Scholz, B. 2005. Incremental map generation with GPS traces. In *Proceedings of IEEE Intelligent Transportation Systems 2005*, pages 574–579. IEEE.

Cao, X., Cong, G., and Jensen, C. S. 2010. Mining significant semantic locations from GPS data. In *Proceedings of the VLDB Endowment*, pages 1009–1020. VLDB Endowment.

Chen, Y. and Krumm, J. 2010. Probabilistic modeling of traffic lanes from GPS traces. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 81–88. ACM.

Cipeluch, B., Jacob, R., Winstanley, A., and Mooney, P. 2010. Comparison of the Accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. In Tate, N. J. and Fisher, P. F., editors, *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, pages 337–340, Leicester, United Kingdom.

Ekpenyong, F., Palmer-Brown, D., and Brimicombe, A. 2009. Extracting road information from recorded GPS data using snap-drift neural network. *Neurocomputing*, 73(1):24–36.

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

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

Haklay, M. and Weber, P. 2008. Openstreetmap: User-generated street maps. *Pervasive Computing, IEEE*, 7(4):12–18.

Heipke, C. 2010. Crowdsourcing geospatial data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6):550–557.

Hochmair, H. 2005. Towards a Classification of Route Selection Criteria for Route Planning Tools. In *Developments in Spatial Data Handling*, pages 481–492. Springer Berlin Heidelberg, Berlin/Heidelberg.

Hochmair, H. H., Zielstra, D., and Neis, P. 2015. Assessing the Completeness of Bicycle Trail and Lane Features in OpenStreetMap for the United States. *Transactions in GIS*, 19(1):63–81.

Kim, B.-K., Jo, J.-B., Kim, J.-R., and Gen, M. 2009. Optimal Route Search in Car Navigation Systems by Multi-objective Genetic Algorithms . *International Journal of Information Systems for Logistics and Management*, 4(2):9–18.

Li, J., Qin, Q., Han, J., Tang, L.-A., and Lei, K. H. 2014. Mining trajectory data and geotagged data in social media for road map inference. *Transactions in GIS*, 19(1):1–18.

Marchal, F., Hackney, J., and Axhausen, K. W. 2005. Efficient map matching of large global positioning system data sets: Tests on speed-monitoring experiment in Zürich. *Transportation Research Record: Journal of the Transportation Research Board*, 1935(1):93–100.

Mooney, P. and Corcoran, P. 2014. Analysis of Interaction and Co-editing Patterns amongst OpenStreetMap Contributors. *Transactions in GIS*, 18(5):633–659.

Ranjitkar, P., Nakatsuji, T., Gurusinghe, G., and Azuta, Y. 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*, pages 608–615. American Society of Civil Engineers, Boston.

Reynolds, D. 2009. Gaussian mixture models. *Encyclopedia of Biometrics*, pages 659–663.

Schroedl, S., Wagstaff, K., Rogers, S., Langley, P., and Wilson, C. 2004. Mining GPS Traces for Map Refinement. *Data Mining and Knowledge Discovery*, 9(1):59–87.

Shoval, N., Kwan, M.-P., Reinau, K. H., and Harder, H. 2014. The shoemaker's son always goes barefoot: Implementations of GPS and other tracking technologies for geographic research. *Geoforum*, 51(C):1–5.

Stopher, P., Clifford, E., Zhang, J., and FitzGerald, C. 2008. Deducing mode and purpose from GPS data. Technical report.

Taylor, M. A., Woolley, J. E., and Zito, R. 2000. Integration of the global positioning system and geographical information systems for traffic congestion studies. *Transportation Research Part C: Emerging Technologies*, 8(1):257–285.

Thierry, B., Chaix, B., and Kestens, Y. 2013. Detecting activity locations from raw GPS data: a novel kernel-based algorithm. *International Journal of Health Geographics*, 12(1):14.

van de Coevering, P., Kroesen, M., Maat, K., and van Wee, B. 2015. Causal effects of built environment characteristics on travel behaviour: a longitudinal approach. In *Proceedings of the Workshop on Activity-Travel Behaviour Dynamics*, pages 1–13, Delft, Netherlands.

Van der Spek, S. C., Van Schaick, J., De Bois, P., and De Haan, A. R. 2009. Sensing Human Activity: GPS Tracking. *Sensors*, 9:3033–3055.

van Winden, K. 2014. Automatically Deriving and Updating Attribute Road Data from Movement Trajectories. Master's thesis, Delft University of Technology, Delft, the Netherlands.

WegenWiki 2014. WegenWiki Rijstrook. `http://www.wegenwiki.nl/Rijstrook`.

Zhang, L., Thiemann, F., and Sester, M. 2010. Integration of GPS traces with road map. In *Proceedings of the Second International Workshop on Computational Transportation Science*, pages 17–22. ACM.

Zheng, Y., Chen, Y., Li, Q., Xie, X., and Ma, W.-Y. 2010. Understanding transportation modes based on GPS data for web applications. *ACM Transactions on the Web (TWEB)*, 4(1):1.

Zheng, Y., Zhang, L., Xie, X., and Ma, W.-Y. 2009. Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th international conference on World wide web*, pages 791–800. ACM.

Zielstra, D. and Zipf, A. 2010. A comparative study of proprietary geodata and volunteered geographic information for Germany. In *Proceedings of the 13th AGILE International Conference on Geographic Information Science*, Guimarães, Portugal.