

The 11th International Conference on Ambient Systems, Networks and Technologies (ANT)
April 6-9, 2020, Warsaw, Poland

Road centreline and lane reconstruction from pervasive GPS tracking on motorways

Mohammad Ali Arman^{a*}, Chris M.J. Tampère^a

^a*Centre for Industrial Management, Traffic and Infrastructure; KU Leuven; Celestijnenlaan 300; 3001 Leuven, Belgium*

Abstract

A lane-based Routable Digital Map is a basis for construction of a floating-car dataset for lane-based traffic analysis. We proposed an algorithm that is capable of identifying lanes in highway segments based on GPS trajectories collected by mobile phones. The algorithm consists of three main steps. First, we identify nodes within the test site, and divide the network into segments. Second, the central line of each segment is identified based on a dissimilarity matrix computed based on Dynamic Time Warping criteria. Finally, the Gaussian Mixture Method is used to identify the lanes. This allows the width of the lanes to remain constant throughout the segment. The results have been validated by comparing the share of traffic volume in each lane based on the trajectory points in the identified lanes and the loop detectors' data. The results show that the proposed algorithm can determine the lanes with acceptable accuracy. Estimating the traffic volume and its speed based on floating car data provides a big step in enabling data fusion from multiple sources more accurately and to estimate traffic state more precisely.

© 2020 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Peer-review under responsibility of the Conference Program Chairs.

Keywords: Lane-based map, Routable digital maps, Lane Identification, Lane-based traffic data.

1. Introduction

With the development of ITS applications the concept of as Routable Digital Maps (RDM) has emerged. These maps are enhanced with features such as travel time, speed limit etc. In the recent years, several researches have focused on the concept of RDMs and tried to enhance them. Using GPS data to develop RDMs is known as Automated Map Construction (AMC). For at least two main reasons it is important to upgrade RDMs to a lane-based level. First which is the main motivation of this paper is preparing a dataset for lane-base traffic analysis based on floating car data (trajectory data). There exist a very few dataset on lane change behaviours and developing lane-based RDMs will

* Corresponding author. Tel.: +32 16 32 16 73.

E-mail address: mohammadali.arman@kuleuven.be

enable us to extract lane change behaviour from trajectories. And second, in the development of on-line GPS-based driver guidance software. Efforts for this upgrade suffer from two main shortages. First, the road centreline that is a basis for identification of the lanes is mostly derived based on the density of trajectories. This can lead to bias in the centreline as a result of change in the distribution of traffic flow over the width of the road, e.g. in or near complex segments such as weaving sections. And second, almost always the benchmark used to measure the results was a reliable map as ground truth, and traffic benchmarks such as conformity of the proportion of vehicles in each lane based on identified lanes and loop detector data were never considered.

The main aim of this paper is to propose an algorithm that is capable of identifying lanes in highway segments based on GPS trajectory tracks collected by mobile phones. We first segmented our highway test network, through identification of nodes. Then the central line of each segment is identified. Finally, the Gaussian Mixture Method (GMM) is used to determine the road lanes. We used the root-mean-square error (RMSE) for validation. The rest of this paper is organized as follows: section two reviews existing literature. We introduce the test site and collected data in section three. Our methodology is explained in section four and the lane identification results as well as validation procedure are presented in section five. Finally, we conclude our study and propose some directions for further studies in the last section of the paper.

2. Literature Review

AMC algorithms have been classified in the literature based on the core algorithm that is used to construct the road centreline. There are some studies that tried to classify AMC algorithms and perform qualitative, quantitative, and comparative evaluations on them [1, 2]. There are lots of studies aimed to construct accurate RDMs but only in road centreline level and without any information about lanes, see for instance [3-5].

One of the earliest attempts for lane identification based on low-precision GPS data was done by Wagstaff *et al.*, using the K-means algorithm for clustering of lateral distance of GPS tracks [6]. Another study used 55 vehicles that were equipped with standard GPS devices and their data was recorded for several days to construct a lane-based RDM. They used perpendicular lines at certain distances relative to the centreline of the road for one-dimensional classification of the trajectories [7]. The main drawback of this study is that it runs the risk of discontinuity of lane width along the road segment. A recent alternative for the previous classification methods for lane identification is a Naive Bayes classifier which claimed it is capable to identify lanes even over low-precision GPS trajectories [8]. The authors of [8] also examined the use of the one-dimensional GMM method in detecting lanes and compared the results with satellite images [9]. Also a fuzzy-set-based algorithm has been proposed for limited applications of lane-level vehicle positioning near signalized intersections and using smartphones' GPS data [10]. Yang *et al.* in their two-step fuzzy logic-based approach, first match the trajectory data with existing maps and then updated these maps based on the detected lanes [11].

3. Data Collection and Test Network

Our test network consists of the E313 and R1 highways between junctions Wommelgem and Antwerpen-Zuid in both driving directions near Antwerp, Belgium. The data was collected through the Touring Mobilis smartphone app of Be-Mobile for both iOS and Android operating systems. Trajectory information including time, coordinates, speed, headings, and a dedicated ID for each vehicle is stored at 1Hz frequency during the entire year 2019. On weekdays, trajectories of on average 4220 vehicles are registered.

4. Methodology

A three-step algorithm that converts a set of trajectory data into a lane-level RDM. Each step uses universal methods that can be replicated with similar data on similar highway networks. Fig. 1 gives an overview of the methods, which are detailed hereafter.

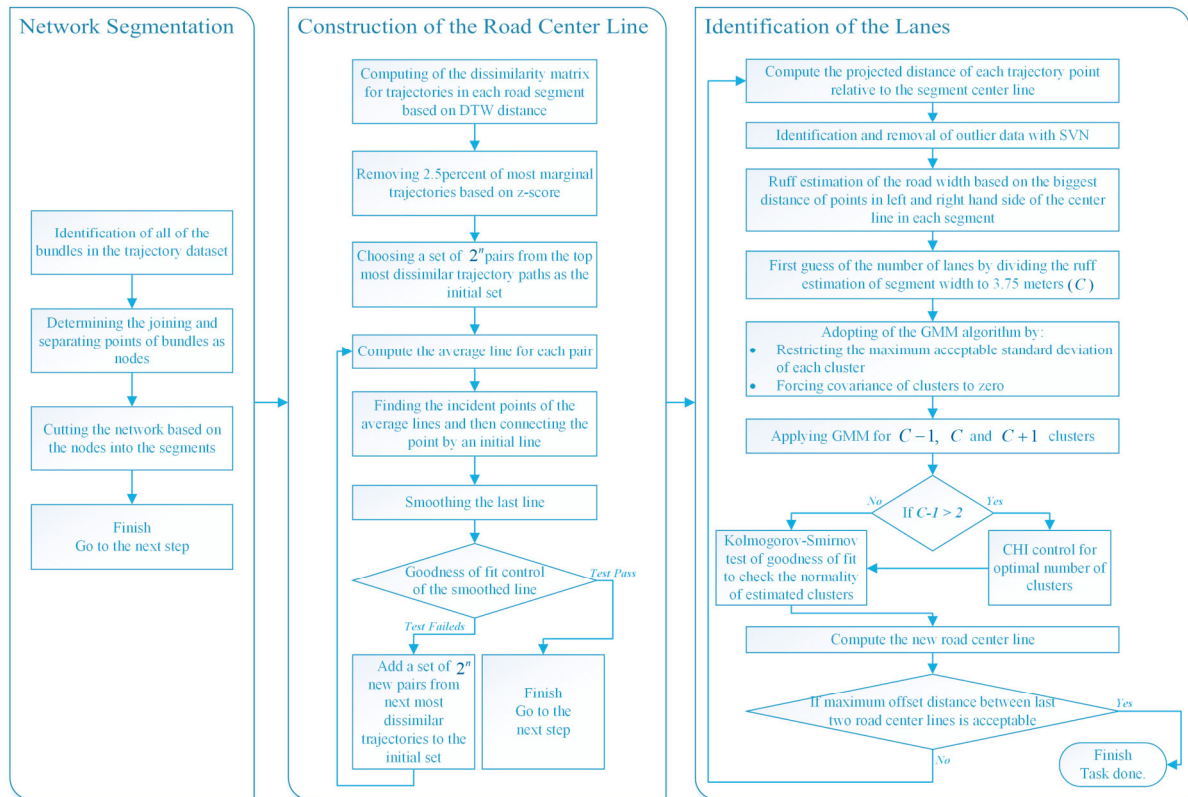


Fig. 1. The flowchart of the proposed algorithm

As the first stage, to simplify the classification process, we identify nodes that divide the test network into homogeneous road segments (in terms of the number of the lanes). Nodes are defined where two bundles of trajectories are joined together (merge) or separate (diverge). A bundle is defined as a set of trajectories in a road segment that follow the same path through the network; bundles are identified using the QuickBundles (QB) algorithm [12]. Fig. 2 shows two situations where bundles merge and diverge. The algorithm considers incidences, i.e. intersections of trajectory paths belonging to different bundles (due to lane changes and noise on the lateral position recordings). Sufficiently upstream of a diverge (resp. downstream of a merge), the number of incidences per distance unit is relatively stable; we call this stable quantity M for a diverge (respectively M' for a merge). The more downstream of a diverge, the more the two bundles are separated and the number of incidents tends from M to zero. Likewise upstream of a merge, the number of incidences tends from zero to M' . In almost all cases in our data, the point where the last (resp. first) incidence occurs is a GPS data error. The points where the number of incidences equals 3.5% of M (resp. 2.7% of M'), appear in our data to be the closest approximation to the points of the diverge (resp. merge) in the geometrical plan; this is where we define the node.

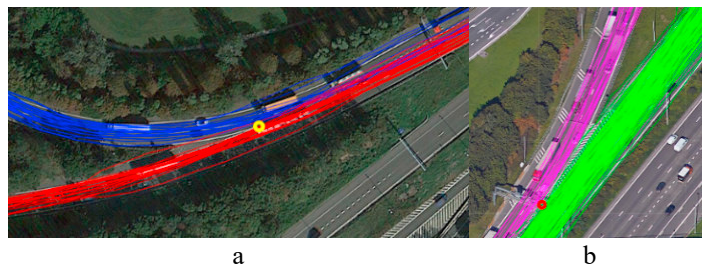


Fig. 2. (a) a node where two bundles of trajectories diverge; (b) a node where two bundles of trajectories merge.

The second stage of our algorithm is the construction of the road centreline. Previous studies have used two main approaches. Some studies take the road centreline from existing RDMs (usually Google Maps or OpenStreetMaps (OSM)), others determine it based on density-based algorithms. [16] suggests that also Google and OSM build their RDMs at least partly based on the density of GPS trajectories (in addition to other sources such as satellite and aerial images). The distribution of density of trajectories over the width of the road is however not constant, as traffic flow distribution over lanes is affected by lateral manoeuvres at complex sections (such as merges, diverges, on/off-ramps, weaving sections). This may explain why, even if we may assume that large quantities of tracks were used, the road centrelines of Google Maps and OSM (shown in Fig. 3) are not consistent: they do not overlap and their relative distance varies along the segment.

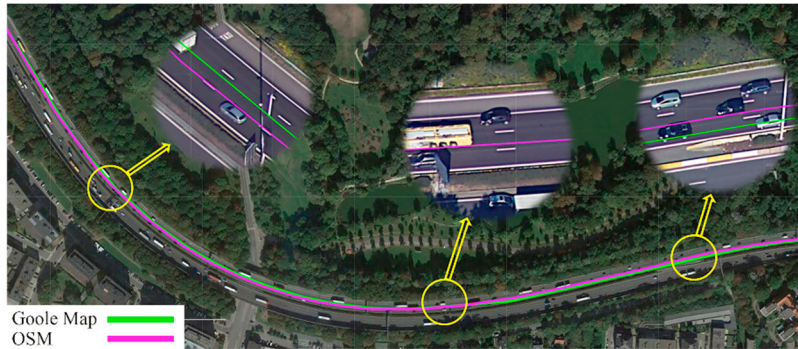


Fig. 3. Comparison of the road centerline in a highway segment near Antwerp which extracted from Google Map and OSM

As an alternative we therefore developed a method based on the distance of the trajectory paths as curves. In this method the distance between all of the trajectory paths which are passing a road segment are computed in a pairwise manner based on the Dynamic Time Warping (DTW) distance D^i between each pair of trajectory paths [13]; then a normalized dissimilarity score $0 < S_{DTW}^i \leq 1$ is computed based on these distances:

$$S_{DTW}^i = \frac{D^i - \min_j(D^j)}{\max_j(D^j) - \min_j(D^j)} \quad (1)$$

Pairs with higher scores are those that are transversely farther apart from each other in a road segment. In order to avoid the use of unrealistic trajectories (e.g. too far off the road centre due to GPS error), we remove outlier trajectories in our dissimilarity matrix based on the z-score. For our data set, a z-value equal to 1.96 was sufficient to exclude trajectories with high error (e.g. projected off the road pavement) from this step of the algorithm.

Consider then the first 2^n pairs with the highest scores in the dissimilarity matrix (in our experience n equal to 2 or 3 yields good results). In Fig. 4 (a) four pairs of trajectories each in their own colour are chosen for illustration. For each pair, an average line is then constructed. Between the nodes identified in the previous step of the algorithm that bound the segment under consideration, all trajectory paths have a similar length. We can thus calculate with increments of 10 meters along the path, the average lateral position of each trajectory pair (Fig. 4 (b)). The corresponding points of the pair are connected to each other. To obtain a smoother curve, the midpoints of these connections are found and connected to yield one estimated centre line per pair of trajectories. After determining the centre lines for all pairs, a new average centre line is constructed by connecting the points of intersection of the pairs' centre lines as illustrated in the Fig. 4 (c). Finally as shown in Fig. 4 (d), too sharp changes in the angle of successive points in the average centre line are substituted with straight lines to make it more smooth. This smoothed average centre line is the candidate centreline of the road segment.

Another test must be done to make a final decision. To perform this test, a set of $m \times 2^n$ (we find that 3 is a good value for m) pairs of trajectories should be selected from the highest scoring pairs (except those previously selected) based on the dissimilarity matrix. These new pairs are only for evaluation of the centreline. Afterward the average line for each pair is determined. In the next step, the projected distance of the new average lines with respect to the candidate line is calculated in every 10 meters as a measure of error. Finally, the Mean Square Error (MSE) test is used to determine if the candidate line is a good estimate of the road centreline or not. We considered a maximum error of 0.1 meter as a criterion for accepting or rejecting a candidate's line based on the MSE results. If the test fails

then we next consider $n_{t+1} = 2 \times n_t$ to repeat the algorithm of construction of the centreline, in other word we increase the number of pairs for construction of the centreline two times of the number of pairs in the previous iteration. Our investigation showed that for almost all the road segments studied in this paper, performing the algorithm once or repeating it for a maximum of two iterations yields a good centreline (as explained with results in section 5.2), in addition the values of n and m obtained only for our test network and may vary for other networks and other data.

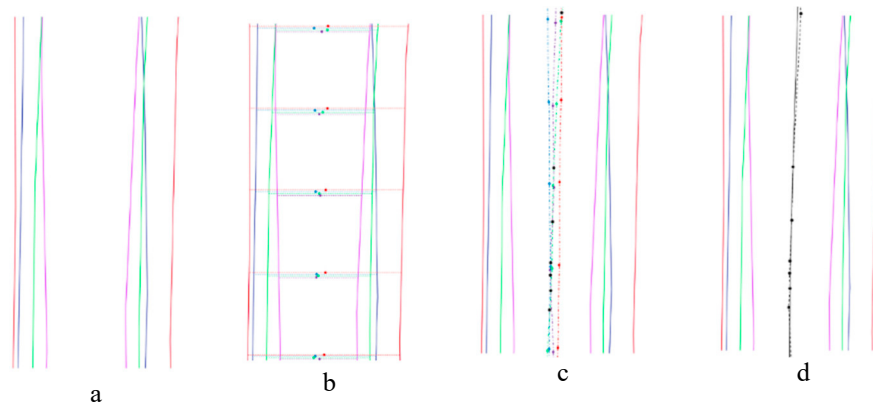


Fig. 4. (a) choosing pairs of trajectories; (b) computing their average distance; (c) determining the average lines; (d) determining the center line.

The third and last stage of the algorithm is the lane identification. In this stage the first action is detecting and removing outliers in trajectories. To this end, the projected distance from the centreline of the road segment is calculated for each trajectory point. Then outlier points based on these distances are identified. We use the one class support vector machine (SVM) method to detect outliers [14]. Then, based on the maximum distance of the trajectory points on the left and right hand sides of the centreline, the maximum possible width of the road segment is obtained. As our test site is a highway network, an initial guess about the number of lanes is given by dividing the maximum width obtained from the previous step by 3.75 meters (let's call this guess C). We used Gaussian Mixture Model (GMM) to identify lanes [15]. The GMM method assumes that the data distribution is obtained from the combination of several normal distributions. At this stage it is impossible to test this assumption and we assume it to be true. But after estimating the clusters (lanes) we will check to what extent this initial assumption was correct. A normal mixed distribution typically consists of a combination of several normal distributions. Each of these normal distributions has their own variance, but their mixture also assumes a covariance between each pair of distributions. For the purpose of this paper, it is essential to assume the normal distributions are independent from each other, hence the covariance is assumed to be zero. Without this assumption, some unrealistic results may be obtained (as we show in section 5.3). Also, given the range of the error in the GPS data, classification of the trajectory data with GMM can lead to the observation of kurtosis in clusters in the sides of the road. We have also noted that since the shares of the vehicles observed in the central lanes are typically larger than other lanes (at least in our test site), due to the density of observations, central clusters (lanes) may be computed too narrow. Consequently, the widths of the side lanes will be larger than 3.75 meters and the widths of the central lanes will be less than the standard. So, we added a constraint to the GMM algorithm regarding the maximum and minimum acceptable standard deviation for clusters. Moreover, an important contribution of this paper is that we classify trajectory data for a whole length of each segment of the road at once, and do not use vertical cross sections. This ensures that the width and position of the lanes remain the same throughout each segment.

As the first guess of the number of the lanes (C) may be rough, we also test GMM for $C - 1$, C and $C + 1$ clusters. We used the Calinski-Harabasz Index (CHI) [16] to determine the optimal value of C , namely C^* . Next, we verify if scattering of the trajectory points on each lane indeed has a normal distribution or not by a Kolmogorov-Smirnov goodness of fit (KSG) test.

Finally, the centreline found in stage 2 is corrected by an offset derived from the identified lanes. For road segments an odd number of lanes, the centreline is shifted towards the centre of the middle lane; for segments with an even number of lanes, the centreline is shifted towards the mathematical intersect point of the identified normal distributions

for two adjacent middle lanes.

If the shifted centreline is offset by no more than a threshold (default: 1cm), the algorithm is finished; otherwise the whole algorithm of identifying of the lanes is repeated based on shifted centreline in a loop until the required shift is below the threshold.

5. Results and Discussion

In this section, the results of the proposed algorithm presented and validation results are described. Five weekday trajectory data from September 2019 was used to identify nodes in our test network. The algorithm found 20 nodes, which corresponds to the actual number of nodes based on the network geometry. Through these nodes, our test network is divided into 34 road segments. The position of the identified nodes was then compared with their geographical location based on Google Earth imageries. The comparison results show that the minimum and maximum distances of the identified nodes are 4.6 and 9.4 m, respectively, relative to their actual geographical location.

The dissimilarity matrix was calculated for all 34 road segments for non-outlier trajectories. The value of n for all road segments is assumed to be 3 (8 pairs). Fig. 5 represents examples of the road centreline for different locations of a four lanes segment in the test network. The significance of the obtained centreline is its consistency throughout its length, even though it is constructed at low cost, solely based on a small amount of smartphone GPS data. One limitation of the resulting centreline is however that it deviates slightly to the curve direction as a result of centrifugal bias of GPS data registered by smartphones (due to their relatively simple internal filtering) especially in high speed, curvy segments.

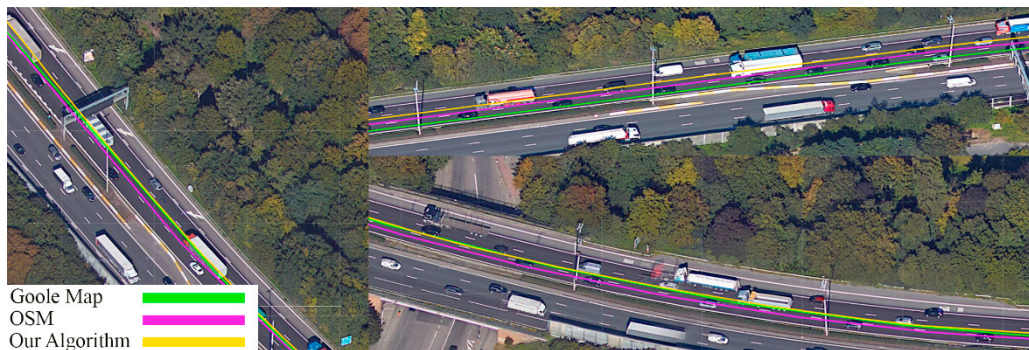


Fig. 5. Calculated road center line in a segment in the test site

As explained earlier, in the clustering of trajectories we assumed that the clusters' covariance is equal to zero. In addition, the standard width of the lanes in the test network is 3.75 meters. Statistically, 95% of observations are in the range of $\mu \pm 1.96\sigma$. Therefore the acceptable standard value for standard deviation is $0.956 = (3.75/2)/1.96$. Given that this standard deviation comprises 95% of the data, $0.956 \pm 0.956 \times 2.5\%$ is considered as the acceptable standard deviation of the clusters in the GMM method. Therefore, this limitation will be applied in the iterations of the GMM algorithm toward the optimal clusters. Fig. 6 shows a comparison of the effect of applying and not applying these assumptions on the results of the GMM algorithm in a four-lane segment of the test network. Sections "a", "b" and "c" are all calculated for the same data in this figure. In section "a", neither of the two assumptions are considered. In section "b" the assumption of independence of clusters is considered, but no restriction on the acceptable minimum and maximum standard deviations is applied. In Section "c", both assumptions are considered. Only applying both assumption could guarantee clustering results truly representing road lanes.

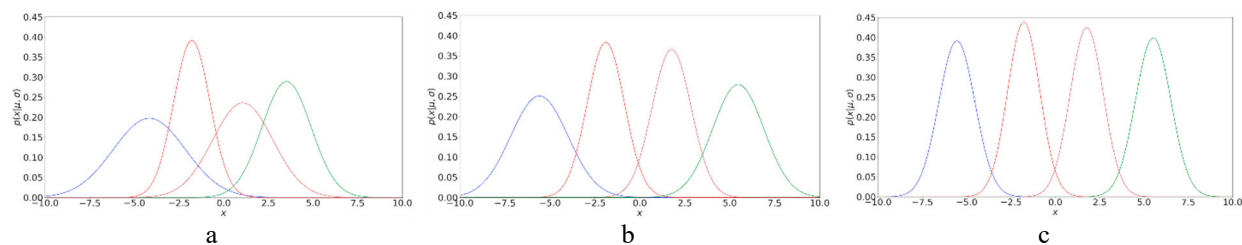
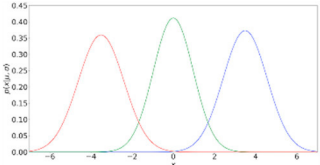
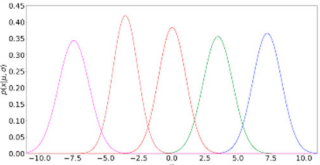
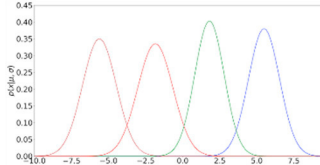


Fig. 6. Comparison of the effect of applying and not applying of the studied assumptions on the results of a GMM algorithm

With detected nodes, the test network was divided into 34 segments. Table 1 presents lane identification results in some segments as examples. Overall, in more than 70% of cases, KSG results were obtained at all significant α levels and estimated clusters follow normal distribution. In other cases, for $\alpha = 0.2$ the hypothesis of the KSG test is rejected but for the other levels of α , the test hypothesis is confirmed.

Table 1. Sample results of the lane identification step

Location in the test network	a			b					c			
Probability distribution of the clusters												
Separation point of clusters	-1.80	1.91		-5.49	-1.80	1.85	5.62		-3.58	0.04	3.71	
Lane width	3.81	3.71	3.91	3.84	3.69	3.65	3.77	3.86	3.89	3.63	3.67	3.84
Results of KSG test	0.0548	0.0407	0.0760*	0.0617	0.0572	0.0350	0.0338	0.0535	0.0596	0.0465	0.0315	0.0408

The most important impetus for this study was to obtain the lane-based traffic data, which leads us validation through traffic-based comparison. We chose as comparison criteria the average speed of each lane and average share of each lane from the total traffic flow both in 15 minutes intervals based on trajectories and loop detectors data, at the location of the loop detectors (hereafter called measurements). We then compute root-mean-square error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (MT_i - MD_i)^2}{n}} \quad (2)$$

In this equation MT and MD are measurement based on trajectories and loop detectors respectively and n is the number of time intervals per location. We performed the validation for two four-hour intervals in the morning peak (from 7:00 to 11:00) and in the evening peak (from 14:00 to 18:00), yielding $n=32$ time periods per loop detector. RMSE has the same unit as the variables so its absolute value is not proper for comparison. We normalized RMSE by a normalization factor NF . For the traffic volume share, we estimate NF as the expected order of magnitude, which in case of homogeneous distribution equals 1 divided by the number of the lanes; for speed NF is simply based on the average of speed per observations:

$$NRMSE = \frac{RMSE}{NF} \times 100 \quad (3)$$

The validation results for the same segments as shown in Table 1 are summarized in Table 2. One main reason for higher error in the traffic volume lane share is the centrifugal deviation of the GPS records in the road curvatures which especially increase with increase of the speed.

Table 2. The validation results

Road Segment	Number of per lane loop detectors	n*	Normalized RMSE	
			Based on per lane average speed	Based on per lane share of traffic flow
a	6	183	3.34%	12.27%
b	10	307	5.21%	16.26%
c	20	622	4.57%	14.18%

* On various days because of some seasons such as maintenance, loop detector data may not be recorded continuously during the study period.

6. Conclusion

In this paper, we present an algorithm for construction of the lane-based RDMs. The main motivation of the paper was to provide a framework for the extraction and establishment of lane-based traffic databases from floating car data (trajectories). The results presented in this paper show that the proposed algorithm can determine the road lanes. In addition, with respect to traffic measurements such as speed and lane share from total traffic flow, we obtain close approximations solely based on trajectory data with an acceptable range of error (in average near to 4% in term of speed and near to 14% in term of lane share from traffic flow). This method provides a clear perspective of the possibility of achieving the ultimate goal of our research. This is a big step in enabling data fusion from multiple sources more accurately and to estimate traffic state more precisely. There are, of course, some considerations and there are some limitations in the current algorithm. For example, lane drop or lane rise in a segment where it is not any detected node by the proposed algorithm can have some impacts on the results. It may also be possible to measure the optimal transversal position of the lanes relative to the whole width of the road cross-section by applying some additional constraints on the GMM. These can be considered as directions for further extension of this research. In addition, this paper does not investigate the minimum number of samples (vehicles) to obtain valid results from the proposed algorithm. Such an analysis could be one of the possible attempts in the future development of this paper.

7. References

- [1] Biagioni, J. and J. Eriksson, *Inferring road maps from global positioning system traces: Survey and comparative evaluation*. Transportation research record, 2012. **2291**(1): p. 61-71.
- [2] Ahmed, M., et al., *A comparison and evaluation of map construction algorithms using vehicle tracking data*. GeoInformatica, 2015. **19**(3): p. 601-632.
- [3] Pollak, K., A. Peled, and S. Hakkert, *Geo-based statistical models for vulnerability prediction of highway network segments*. ISPRS International Journal of Geo-Information, 2014. **3**(2): p. 619-637.
- [4] Xie, X., et al., *Detecting road intersections from GPS traces using longest common subsequence algorithm*. ISPRS International Journal of Geo-Information, 2017. **6**(1): p. 1.
- [5] Tang, J., et al., *An Automatic Method for Detection and Update of Additive Changes in Road Network with GPS Trajectory Data*. ISPRS International Journal of Geo-Information, 2019. **8**(9): p. 411.
- [6] Wagstaff, K., et al. *Constrained k-means clustering with background knowledge*. in *Icml*. 2001.
- [7] Chen, Y. and J. Krumm. *Probabilistic modeling of traffic lanes from GPS traces*. in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 2010. ACM.
- [8] Tang, L., et al., *Lane-level road information mining from vehicle GPS trajectories based on naïve bayesian classification*. ISPRS International Journal of Geo-Information, 2015. **4**(4): p. 2660-2680.
- [9] Tang, L., et al., *CLRIC: collecting lane-based road information via crowdsourcing*. IEEE Transactions on Intelligent Transportation Systems, 2016. **17**(9): p. 2552-2562.
- [10] Marinelli, M., et al., *A Fuzzy set-based method to identify the car position in a road lane at intersections by smartphone GPS data*. Transportation Research Procedia, 2017. **27**: p. 444-451.
- [11] Yang, X., et al., *Automatic change detection in lane-level road networks using GPS trajectories*. International Journal of Geographical Information Science, 2018. **32**(3): p. 601-621.
- [12] Garyfallidis, E., et al., *Quickbundles, a method for tractography simplification*. Frontiers in neuroscience, 2012. **6**: p. 175.
- [13] Müller, M., *Dynamic time warping*. Information retrieval for music and motion, 2007: p. 69-84.
- [14] Castillo, E., et al., *Distributed One-Class Support Vector Machine*. International Journal of Neural Systems, 2015. **25**(07): p. 1550029.
- [15] Fraley, C. and A.E. Raftery, *How many clusters? Which clustering method? Answers via model-based cluster analysis*. The computer journal, 1998. **41**(8): p. 578-588.
- [16] Desgraupes, B., *Clustering indices*. University of Paris Ouest-Lab Modal'X, 2013. **1**: p. 34.