Clustering-based methodology for estimating bicycle accumulation levels on signalized links

A case study from the Netherlands

Reggiani, Giulia; Dabiri, Azita; Daamen, Winnie; Hoogendoorn, Serge

DOI
10.1109/ITSC.2019.8917138

Publication date
2019

Document Version
Accepted author manuscript

Published in
2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019

Clustering-based methodology for estimating bicycle accumulation levels on signalized links: a case study from the Netherlands

Giulia Reggiani, Azita Dabiri, Winnie Daamen, Serge Hoogendoorn

***Abstract*— The number of queued bicycles on a signalised link is crucial information for the adoption of intelligent transport systems, aiming at a better management of cyclists in cities. An unsupervised machine learning methodology is deployed to produce estimations of accumulation levels based on data retrieved from a bicycle street of the Netherlands. The use of a clustering-based approach, combined with a conceptual insight into the bicycle accumulation process and various data sources, makes the applied methodology less dependent on sensor errors. This clustering-based methodology is a first step in bicycle accumulation estimation and clearly identifies levels of cyclists accumulated in front of a traffic light.**

I. INTRODUCTION

There is an evident increase of bicycles trips in cities [8]. This leads, among other things, to long waiting times at traffic lights and unsafe situations such that municipalities, in some countries like the Netherlands, are struggling to manage bicycle traffic. Intelligent Transportation Systems (ITS) can mitigate the situation, as already proven effective in vehicular transport management, by e.g. 1) reducing delay using adaptive traffic signal controllers [9], or 2) reducing discomfort by providing real time traffic information [5], valuable for user's route choice. In order to deploy such ITS, real-time and accurate information about bicycle accumulation on urban cycle paths is crucial.

The fundamental difference between car and bike queues, depends on the unstructured and non lane based behaviour of cyclists. Consequently, fixed location sensors incorporate counting errors which have an effect on flow data of bicycles, leading to growing cumulative errors while estimating accumulation. In vehicle queue estimation studies, the cumulative error problem is a well-known research topic investigated in [3] and [11], to name but a few. To the best of the authors' knowledge, the cumulative error problem has not been addressed yet in the bicycle domain.

We propose a clustering-based methodology for bicycle accumulation estimation, applicable to various kinds of unlabeled and error prone data. Although, to the best of the authors' knowledge clustering has never been applied to bike accumulation problem, for an overview of its applications in transport domain see [2], [12]. The use of a clustering-based method combined with a conceptual insight into the bicycle accumulation process and various data sources make the applied methodology less dependent on sensor errors. The methodology is tested on field data retrieved from inductive loop sensors and represents the first step in setting up a methodology for bike accumulation estimates.

Supervised machine learning techniques require a large amount of labels (i.e. ground truth), not always available, to train the models. Instead, the low data requirements of unsupervised machine learning techniques make this methodology attractive and easy to implement for practitioners. The achieved bicycle accumulation levels can be used as real time traffic level indicators of edges of a cycle network or as input for data driven control measures.

This article is organized as follows. Section II presents the methodology (i.e. the general research approach). Whereas, section III illustrates, through a case study, the methods used within the methodology and evaluates the proposed clustering-based estimation approach, by comparing it with other estimation techniques. Finally the conclusions are reported in section IV.

II. RESEARCH METHODOLOGY

Fig.1 shows, on the left side, the general unsupervised approach for researchers that aim to estimate bicycle accumulation on a signalized link. It consists of seven steps, each is explained in the following subsections. The grey diagram, on the right in Fig.1, illustrates the steps performed in the case study of section III. Practitioners addressing scenarios similar to the one of the case study (i.e. same data availability and setting) can estimate bicycle accumulation by applying steps 4, 5 and 6 of the methodology.

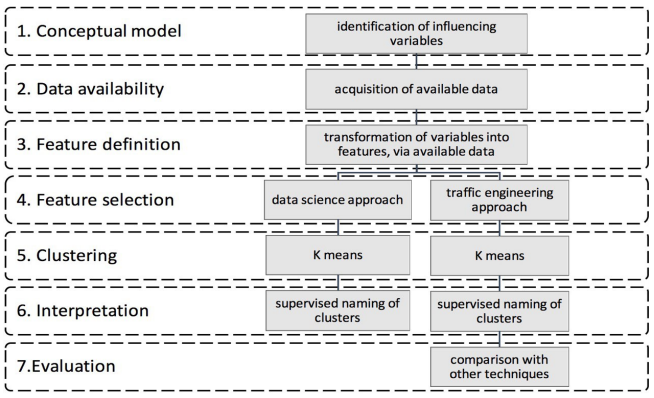


Fig. 1: Steps of the research methodology

*This research was supported by the ALLEGRO project, which is financed by the European Research Council 40 (Grant Agreement No. 669792) and the Amsterdam Institute for Advanced Metropolitan Solutions.

G. Reggiani, A. Dabiri, W. Daamen and S. Hoogendoorn are with department of Transport and Planning, Technical University of Delft, Delft, The Netherlands. email: g.reggiani@tudelft.nl, a.dabiri@tudelft.nl, w.daamen@tudelft.nl, s.p.hoogendoorn@tudelft.nl

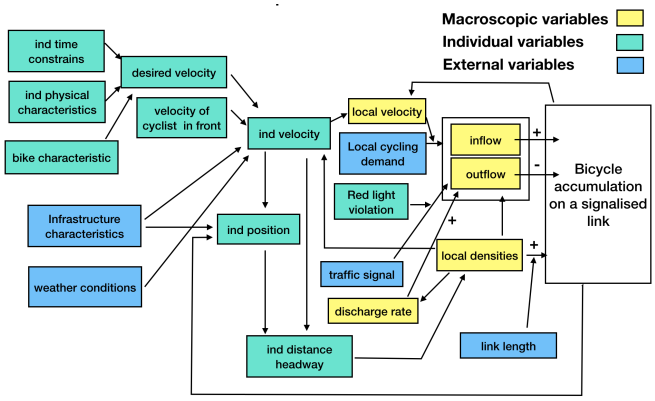


Fig. 2: Conceptual model for bicycle accumulation

A. Conceptual model

The conceptual model, as shown in Fig. 2, highlights the influencing factors for bicycle accumulation and their mutual relationships. This conceptual model is used to 1) understand which are the influential factors that can be included as features in the learning process 2) avoid using available data without motivated correlation to bicycle accumulation 3) improve the estimation model by identifying which additional variables to measure in future studies. Only dominant relationships are included in the figure.

The conceptual model shows, above all, that:

- macroscopic variables have a direct effect on bicycle accumulation, while individual and external variables have an indirect effect, since they influence some macroscopic variables that in turn determine bicycle accumulation.
- local density, depending on its location (either the cyclist's position or the sensor location), can influence the outflow, individual velocity and queue discharge rate. If local density can be used to approximate average density (i.e. number of accumulated cyclists) then it can also be seen as a direct influencer of bicycle accumulation.
- inflow and outflow have a direct influence on bicycle accumulation. Theoretically, inflow and outflow could fully determine the accumulation through the bicycle conservation law.

B. Data Availability

This step seeks for data sources that can measure influential variables of the conceptual model. Data availability in general depends on the type of sensors road authorities have deployed. We provide an insight into specific data availability based on inductive loop sensors, which are used the case study of section III.

Theoretically, inductive road sensors provide bike counts, occupancy of the sensor over time, and speeds (from two loops close to each other). Having bicycle counts upstream and downstream (i.e. inflow and outflow) we can define bicycle accumulation, through the bicycle conservation law. However, as shown in section III-F, using this conservation law based on bicycle loop sensor signals leads to a cumulative error due to inaccuracies in the downstream loop counts[1]. As a result, we need additional information either continuously or at specific moments to correct for the accumulated error. Thanks to the conceptual model, it is possible to see what other variables to extract from loop sensors to compensate flow errors while estimating bicycle accumulation on a link.

C. Feature Definition

In this step, the output from the previous two steps are combined in order to translate the variables of the conceptual model into features, defined for each observation period δt. An observation period is represented by a vector made of features, that represent the numerical value of one or more independent variables during the observation period. The bicycle accumulation, to be estimated, is the number of bikes, on the link, at the end of each observation period.

D. Feature selection

Feature selection is needed in order to remove irrelevant or redundant variables. Practical experience with machine learning shows that reducing features can improve learning performance by increasing learning accuracy, lowering computational cost, and improving interpretability [1]. The improved performance of clustering with less variables is also confirmed in our case study (see Table II).

Features can be selected by means of a pure data science approach or by means of traffic engineering domain knowledge. The former approach looks at correlation between features and between features and a sample of ground truth accumulation, whereas the latter selects features that are more meaningful from a theoretical point of view, based on the conceptual model.

E. Clustering

This is the core step of the methodology that learns latent patterns within the data without being trained on the corresponding ground truth (i.e. labels). Unsupervised approaches, such as clustering, are preferred to supervised due to the fact that labels are not easily determined for bicycle accumulation level. If the results are to be sufficiently reliable for training purposes, such information needs to be manually extracted from video footage, through a very time-consuming process. By making use of a clustering technique, less ground truth data needs to be extracted because labels are not needed for the training but are only used for the interpretation, as well as the evaluation, of the results.

F. Interpretation

To interpret the latent clusters found in the previous step, a limited amount of ground truth is used. The interpretation step is used to 1) understand if the groups found by the

[1]Whilst cyclists predominantly pass over the upstream sensor one at a time, they tend to cycle over the stop-line sensor (when the traffic light turns green) very close to each other, with a slight sidewise shift.

clustering algorithm have a physical meaning from bicycle accumulation point of view and 2) give names to the classes, according to the accumulation level they represent.

G. Evaluation

For the evaluation of the methodology, other estimation techniques should be used to assess how the clustering methodology estimates compared to other methods. In this paper, the comparison is made with methods based on conservation of bicycle law and a corrected version of it.

III. Case Study

In this section, we use real data to show how unsupervised machine learning can be applied for estimating bicycle accumulation. The structure of this section follows the same steps of the methodology (section II), starting from data availability. The aim of the case study is twofold: 1) test if a clustering methodology achieves accurate estimations of bicycle accumulation and 2) determine if incorporating transport domain knowledge in the feature selection process improves the estimations.

A. Data availability

The most common sensor technology on cycle paths in the Netherlands is inductive loop sensors, usually in a configuration as shown in Fig. 3. Two sensors are installed, one 20 meters at the upstream of the traffic light and one downstream, at the stop-line. Currently this technology is simply used to request a green light if at least one bike is on the link, but it does not count the amount of bicycles.

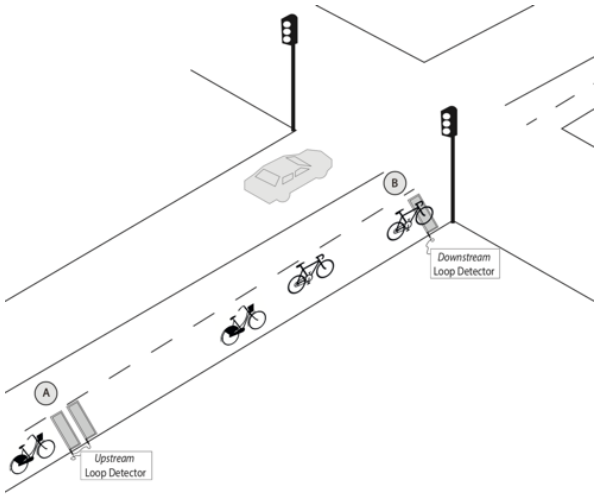


Fig. 3: Loop sensors configuration on cycle paths

The location used as a case study is a signalized intersection, located in the city of Utrecht, the Netherlands, which has very busy morning peak hours. The upstream loop sensor reported over one thousand bikes per hour, during the morning peak. The data from loop sensors were collected over 5 working days, from Monday to Friday, for 4 morning hours, from 6.30 a.m. to 10.30 a.m., and 4 afternoon hours, from 2.30 p.m. to 6.30 p.m.. Overall, 38 hours of loop sensor data (i.e. on-off continuous signal) and its corresponding traffic light signal and camera footage of a 4-hour sample, were made available from the municipality of Utrecht.

B. Feature Definition

Based on the data availability of inductive loop sensors and the insight on the first order influencial variables (from the conceptual model), features that carry information on inflow, outflow, local density, traffic signal and speed have been defined. In order to deploy the estimations in real-time applications, we chose δt to be 30 seconds. In total 14 features are defined for each observation period. Table I lists all the defined features and their corresponding variable of the conceptual model.

TABLE I: list of defined features and corresponding variables

N.	Defined Feature	Conceptual model variable
1	Occupancy Up	local density
2	Occupancy Down	local density
3	Bike Passes Up	inflow
4	Bike Passes Down	outflow
5	Red Occupancy Down	traffic signal & local density
6	Green Occupancy Down	traffic signal & local density
7	Switch	traffic signal
8	Transit	traffic signal
9	Occupancy Up - Green Occupancy Down	traffic signal & local densities
10	Bike Passes Up - Bike passes Down	inflow & outflow
11	Occupancy Up / Occupancy Down	local densities
12	Sum of Speeds	local speed
13	Sum of Occupancies	approx. avg density
14	Occupancy Up - Occupancy Down	local densities

Hereafter, we explain how the fatures in Table I were obtained:

- **Occupancy** is the percentage of time, within one observation period, that the loop sensor is occupied by a bike passing or standing on top of it.
- **Bike passes** measures the number of times the sensor signal has changed within one observation. This is a lower bound estimation of the number of bikes that have passed (i.e. bicycle flow).
- **Red or Green Occupancy** represents occupancy information when the traffic light is either red or green. These features are defined only for the downstream loop because cyclists stop on top of it when the traffic light is red. This results in a high occupancy level downstream, during a red light, that does not reflect a high bicycle density, whereas high occupancy level during the green signal most probably reflects high bicycle density.
- **Switch** and **Transit** features carry information on the traffic light signal. Switch is a percentage that gives information on when, during an observation period, the traffic light last switched from green to red or vice versa. Transit is a binary feature indicating the traffic light state, at the end of each observation period.
- **Speed** feature is a representative approximation of velocities at the sensor location. Speed can be derived from the ratio between flow and density [4]. We approximate flow with bike passes and density with occupancy.

To visualize the relation of each defined feature with the true accumulation of bikes on a signalized link we report

scatter plots with 120 observation periods from one morning hour, for which the ground truth has been manually extracted. Above each scatter plot we report the correlation coefficient between the two variables.

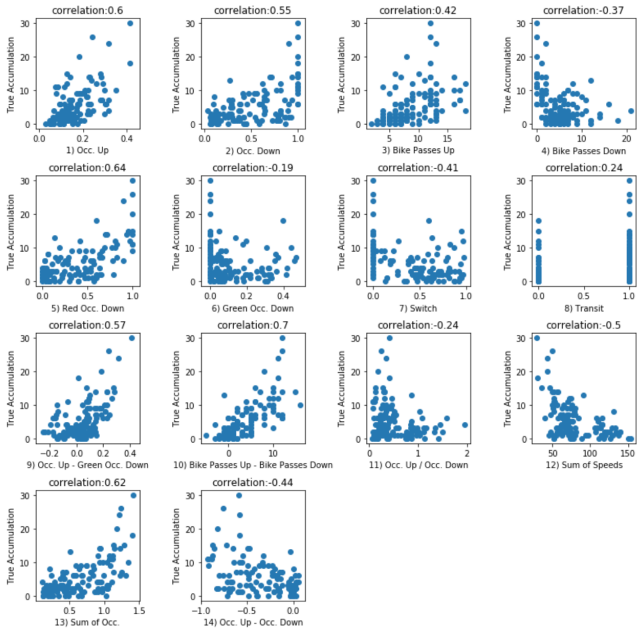


Fig. 4: Relation between features and bicycle accumulation

As expected, Bike passes Up (approximation of inflow) has positive correlation to bike accumulation and Bike passes Down (approximation of outflow) a negative one. Upstream and downstream occupancy and downstream occupancy during red light have a positive correlation, since occupancy is an approximation of local densities. Moreover, features 9 and 10, which respectively represent the difference between upstream and downstream densities and the difference between the inflow and outflow have a strong positive correlation to the bicycle accumulation. This correlation overview will serve as a starting point for feature selection.

C. Feature Selection

From our observations, applying a clustering algorithm directly over all the 14 features does not lead to satisfying results, and the algorithm is not able to cluster together points with similar accumulation. A selection is needed to improve the unsupervised estimation. We propose and implement two different feature selection approaches: 1) a pure data science approach and 2) a domain knowledge approach. The aim is to understand if data driven techniques select the same features as would traffic engineers and if their selections perform differently.

Following a data science approach, we have defined some thresholds, based on observed correlations to bicycle accumulation and among features. This approach first selects the features with a high correlation to the dependent variable, by means of a threshold $T_1 \in \{0.3, 0.4, 0.5, 0.6\}$. Then it drops out all the redundant features by excluding the ones that have a high correlation, with other selected independent variables, by means of a second threshold $T2 \in \{0.5, 0.7, 0.9\}$. Combinations of selected features are ranked based on the average silhouette value S [10]. Top ranking combinations are reported in Table II. The silhouette value is calculated as:

$$S = \sum_{i=1}^{m} s(i) \tag{1}$$

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]} \tag{2}$$

Where:

- $a(i)$ is the average distance between point i and all points within its same cluster
- $b(i)$ is the smallest average distance of point i to all points in any other cluster
- m is the number of all points in all clusters

The silhouette coefficient gives an indication of how similar a point (i.e. an observation period) is to points within its own cluster and how different it is compared to points in other clusters. Hence, this coefficient gives a quality score on how well each observation period has been clustered, given the selected features. The silhouette value is also used to find the best number of clusters, given the selected features.

TABLE II: feature combinations selected based on pure data science approach (only combinations with $S \geq 0.4$ are reported)

	silhouette value	number of clusters	Occupancy Up	Occupancy Down	Bike Passes Up	Bike Passes Down	Red Occupancy Down	Green Occupancy Down	Switch	Transit	Occupancy Up - Green Occupancy Down	Bike Passes Up – Bike passes Down	Occupancy Up / Down	Sum of Speeds	Sum of Occupancies	Occupancy Up - Down
Feature selections	0.44	5							✓		✓					✓
	0.60	3									✓				✓	
	0.45	2	✓								✓	✓		✓	✓	
	0.72	2													✓	
	0.60	2										✓			✓	
	0.40	2	✓	✓					✓		✓					
	0.40	2	✓	✓	✓				✓		✓	✓		✓		
	0.52	3	✓	✓							✓					
	0.45	2	✓	✓							✓	✓		✓		
	0.77	2					✓									
	0.63	2					✓					✓				

Alternatively, from a domain knowledge perspective, we look at the conceptual model (Fig. 2). The model shows that inflow and outflow have a direct influence on bike accumulation, for this reason the difference in upstream and downstream bicycle counts (feature 10) is selected. As seen in Fig. 2, local densities can influence several variables including bicycle accumulation. As a feature to represent densities, feature number 9 is selected because it incorporates the information of both up and downstream densities and also part of the traffic signal downstream. In particular, feature 9 defines the difference between the occupancy upstream and occupancy downstream during a green traffic light. Including more correlated features did not improve clustering performances, thus, from now on, when referring to the

features selected through a knowledge-base approach we refer to features 9 and 10.

D. Clustering

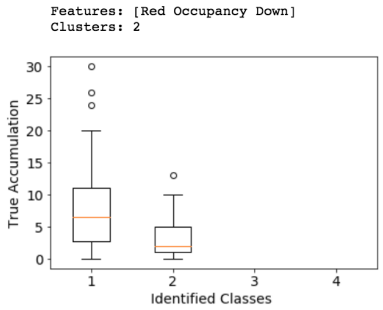
As unsupervised methodology, we apply k-means clustering. This is the starting point for exploring unsupervised machine learning in many domains due to its simplicity and low computation time [7]. K-means algorithm finds a predetermined number of clusters in a dataset. Data are grouped into clusters by minimising the distance between each point and the mean (i.e. centroid) of the assigned cluster. Many distance functions can be used; however, it is common practice for the k-means to use the euclidean distance between points and the centroid. The number of clusters is decided based on the highest silhouette value for a given feature data set. The data science approach, tests 2,3,4,5 number of clusters and selects the one that returned the highest silhouette value.

E. Interpretation

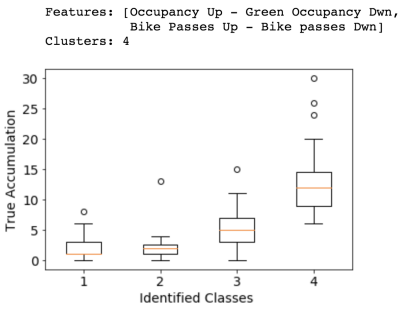
We interpret the unsupervised learning methodology by using ground truth of 120 observation periods of one morning hour. This sample of ground truth is an unbiased representative of our dataset, because it includes a wide range of accumulation values (0 to 30 accumulated cyclists) and not only low accumulation values, as it is the case for other hours within a day.

The interpretation step uses box plots showing bicycle accumulation values, of a sub sample of the dataset, contained in each identified cluster. Fig. 5 reports box plots for the best feature combination (i.e. highest S) of the data-approach and knowledge-approach. The best feature combination based on the data-approach results to be one feature: Red occupancy Down. Clustering performed on this feature vectors results in two classes which have different mean values, but overlap with respect to the amount of cyclists, as shown in Fig. 5a. Whereas, the knowledge-based approach clusters observations into 4 classes (which we name $very\ low, low, medium$ and $high$ accumulation) with lower overlap in values (Fig. 5b).

In general, four classes are more valuable than two, from a traffic engineering perspective. The knowledge-based feature selection clearly distinguishes the high and medium bicycle accumulation class, from other classes (i.e. there is low overlap between accumulation in the high and medium classes). The low and very-low accumulation class represent similar true accumulation levels. However, from a traffic engineering perspective it is more crucial to have an accurate estimation of the high and the medium accumulation class, compared to the low and very low accumulation. As a fact, it is the high accumulation levels that require real time ITS solutions. Comparing the figures, it can be inferred that the better feature combination is the one selected with a knowledge-base approach.The data driven feature selection does not perform as well as the domain knowledge selection. The reason behind such discrepancy requires further investigation but one reasoning may be that the data driven approach selects features that separate points in space better than the features selected with domain knowledge. However those points are separated into clusters which do not relate to bike accumulation but some other, less interpretable, traffic variable. In the following, we proceed by comparing the clustering results from the knowledge-based approach with other accumulation estimation methods.



(a) Data-based feature selection



(b) Knowledge-based feature selection

Fig. 5: Interpretation of the 2 feature selection approaches

F. Evaluation

This section shows how estimation based on the proposed clustering methodology performs compared to the two following approaches:

- Estimation without correction, which is based on the conservation of bicycle law:
  $Acc(i) = Acc(i-1) + Inflow(i) - Outflow(i)$
  where i is the i-th 30-sec period and $Acc(i)$ is the accumulation of bikes at the end of the i-th period.
- Benchmark estimation, which is based on the assumption that all the accumulated cyclists discharge within the first green light phase they encounter
  $Acc(i) = Inflow(i) - Outflow(i)$

This last assumption is based on empirical observations indicating that all accumulated cyclists discharge within the next green light. This means that if we compute estimations every traffic light cycle (start of the green phase), then the accumulation can be calculated as inflow minus outflow. It seems reasonable to extend this assumption to estimate accumulation over fixed time intervals of 30 seconds.

To evaluate the methodology, clustering estimations are represented by the mean accumulation value of the cluster

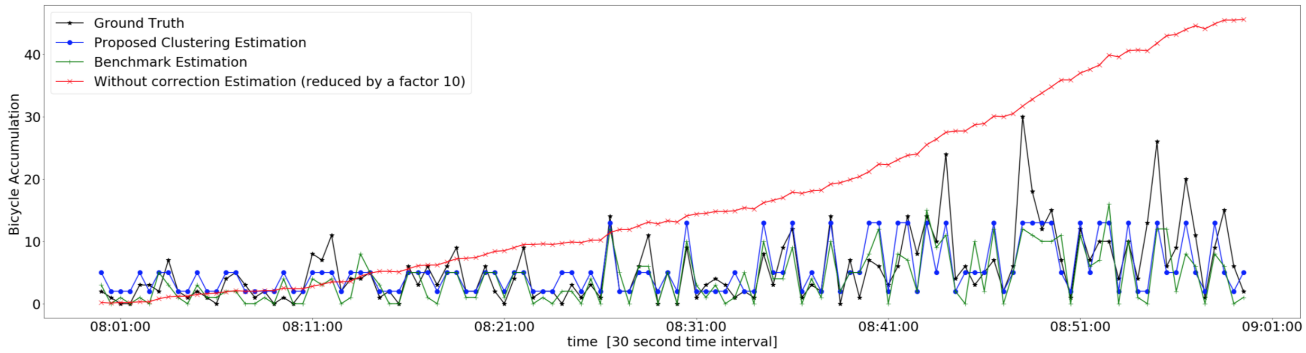


Fig. 6: Evaluation of clustering-based estimation: comparison of 3 accumulation estimation methods

they are part of. Comparison of the methods in Fig. 6 indicates that unsupervised learning methodology has two advantages: it avoids the cumulative error problem and it reduces the means square estimation error compared to the benchmarking method. From Fig. 6 it is evident that estimation based on conservation of bicycle law, leads to a huge cumulative error due to the inaccuracies in the downstream loop counts. Fig. 6 clearly illustrates how closely the estimates from the other two methods, benchmark estimation and clustering-based, follow the ground truth. If we compare these two methodologies based on the mean square error (MSE), the clustering methodology overall performs better, by having MSE=13.75 compared to MSE=17.64 resulted from the benchmark estimation approach.

IV. Conclusions

This work proposes a methodology to estimate the bicycle accumulation levels on a signalized link by using an unsupervised learning technique. The estimation of accumulation levels based on this unsupervised learning method does not require large amounts of ground truth to train the model. Field testing of the methodology on real data indicates an accurate estimation performance with low data requirements, which make it an easily applicable estimation method.

Results from the interpretation step show that incorporating traffic domain knowledge is important to select features for clustering. Instead, applying a pure data driven feature selection, based on correlation values and silhouette coefficient did not show valuable clustering of the data.

Thanks to loop sensor data, which is largely available in urban cycle paths of the Netherlands, a clustering technique can identify the levels of bicycle accumulation. Such information can be used as real time traffic information over an urban network or transmitted to traffic responsive signal controlers, in order to optimally determine green and red light phase of the signal. Clustering can only estimate levels of accumulation and not the exact amount of accumulated bikes. However this characteristic makes the method suitable to be used in many data driven control measures that, for computation reasons, do not require exact but average accumulation levels [6].

Future works should assess how the difference between car and bike data, might affect performance of the method if applied to car queue estimation. Moreover the number of clusters and their interpretation highly depends on the amount and quality of the ground truth data. To overcome this limitation, it is recomended to apply and evaluate the methodology to more than one intersection, with different traffic pattern. Finaly, it is of interest to consider different unsupervised techniques, such as the fuzzy clustering to overcome some of the limitations of the k-means method. So far, our focus is on using largely available data (such as inductive loop sensors) without the need of deploying new or more sophisticated bicycle sensor on the roads. However, the proposed unsupervised approach can be applied in different settings and with various data sources.

References

[1] Salem Alelyani, Jiliang Tang, and Huan Liu. Feature selection for clustering: A review. In *Data Clustering: Algorithms and Applications*, 2013.

[2] Mehmet Ali Silgu and Hilmi Berk Çelikoğlu. K-means clustering method to classify freeway traffic flow patterns. *Pamukkale University Journal of Engineering Sciences*, 20:232–239, 06 2014.

[3] Zahra Amini, Ramtin Pedarsani, Alexander Skabardonis, and Pravin Varaiya. Queue-length estimation using real-Time traffic data. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, pages 1476–1481, 2016.

[4] Patrick Athol. Interdependence of Certain Operational Characteristics within a Moving Traffic Stream. *Highway Research Record*, 72:58–87, 1965.

[5] Giselle de Moraes Ramos, Emma Freijnger, Winnie Daamen, and Serge Hoogendoorn. Commuters and traffic information: A revealed preference study on route choice behavior. In *TRAIL Beta-Congress: Mobility and logistics -Science meets practice, Rotterdam, The Netherlands, 30-31 Oct. 2012*, number October, 2012.

[6] Samah Mohamed El-shafie Mohamed El-tantawy. *Multi-Agent Reinforcement Learning for Integrated Network of Adaptive Traffic Signal Controllers*. PhD thesis, University of Toronto, 2012.

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2009.

[8] ITF. Cycling, Health and Safety. Technical report, 2013.

[9] Tung Le, Péter Kovács, Neil Walton, Hai L. Vu, Lachlan L.H. Andrew, and Serge S.P. Hoogendoorn. Decentralized signal control for urban road networks. *Transportation Research Part C: Emerging Technologies*, 58:431–450, 2015.

[10] Peter J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 29(20):53–65, 1987.

[11] Paul van Erp, Victor L Knoop, and Serge P Hoogendoorn. Estimating the vehicle accumulation : Data-fusion of loop-detector flow and floating car speed data. In *97thTransportation Research Board Annual Meeting*, 2017.

[12] Wendy Weijermars. *Analysis of urban traffic patterns using clustering*. 2007.