Journal of Theoretical and Applied Information Technology
31st July 2020. Vol.98. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN EFFICIENT TRAFFIC STATE ESTIMATION MODEL BASED ON FUZZY C-MEAN CLUSTERING AND MDL USING FCD

FATEMEH AHANIN, NORWATI MUSTAPHA, NASIR BIN SULAIMAN, MASLINA ZOLKEPLI

Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Jalan Universiti 1 Serdang, 43400 Seri Kembangan, Selangor, Malaysia

E-mail: fatemeh.ahanin@gmail.com

ABSTRACT

Monitoring and estimating of large-scale traffic have major role in traffic congestion reduction. Floating Car Data (FCD) is one of the best methods for collecting traffic data due to its versatility and cost efficiency. However, FCD suffers from data sparseness and many researches have been done to improve traffic estimation accuracy with respect to data sparsity. In this paper, a new model based on Fuzzy C-Mean (FCM) clustering and Minimum Description Length (MDL) is proposed to estimate the missing traffic state using FCD. First the Fuzzy clustering is implemented to cluster the road segments based on similarity of their speed at each time slot. Then the MDL principle is applied to estimate the missing traffic state. The experimentation results show that the proposed model can estimate the missing data more accurately than the HMM-based model using the same dataset.

Keywords: *Traffic State Estimation, Fuzzy c-mean Clustering, Pattern Mining, Minimum Description Length, FCD*

1. INTRODUCTION

In today's world, one of the major problems in big cities is traffic management. Monitoring and estimating traffic have major role in traffic congestion reduction. Due to pervasive technologies in telecommunication and transportation systems, there are massive amount of traffic data available. Video cameras, inductive loop detectors and other static sensors can be installed at fixed locations on roads in order to detect traffic state (e.g., flow velocity and traffic density). However, it is difficult for these traditional approaches to cover all roads because they involve extensive infrastructure deployment and high maintenance costs. On the other hand, Floating Car Data (FCD) is a convenient and cost-effective method to gather traffic condition information. It doesn't need any specific device and offers good coverage across road networks with defined penetration rate. However, GPS signals are prone to errors because of urban canyons and tall buildings which affected the traffic estimation results [1]. Existence of these errors has made the estimation and prediction tasks even more sophisticated. Traffic management applications, trip recommendations and any other applications require accurate data to have precise results which are near to actual data. Traffic data includes several traffic states variables (i.e., flow, density, speed and other equivalent variables). As some of the traffic states are missing due to low sampling rate and noise removal, it's necessary to come up with a solution to overcome these issues.

Machine learning offers various methods and models which aid to discover knowledge, recognize hidden pattern, estimate and predict traffic state variables. These methods and models aid to overcome data sparseness and GPS errors which are common issues in many of the gathered data and help to provide more accurate and efficient traffic estimation.

Traffic State Estimation (TSE) refers to the process of the inference of traffic state variables such as flow, density, speed etc. on road segments using partially observed traffic data [2]. Several TSE methods exist which can estimate the missing states. However, there is always a difference between the actual state and the estimated state. This paper aims to propose an efficient model to reduce the estimation error.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

There are two categories for data which been used for TSE namely, single-source data and multi-source data. Single-source data uses one source of data whereas multi-source uses 2 or more sources for data (e.g., FCD with loop detector or video camera.). This paper proposes a model to improve TSE with respect to the data sparsity issue for single-source data. The hypothesis is that by applying the proposed model, the estimation become more accurate and the estimation error will decrease. The proposed model is compared with the HMM-based model [30] and the implementation results indicate that the estimation error is reduced noticeably. The rest of the paper is organized as follows. Section 2 reviews the related work on TSE. Section 3 represent the proposed model based on Fuzzy C-mean and MDL. Section 4 discusses the dataset, parameter tuning and experimentation result. Section 5 concludes the paper.

2. RELATED WORK

This paper has investigated traffic state estimation in the context of FCD and the state of the art of Traffic State Estimation (TSE) methods will be discussed. Recently, working on the FCD has progressed actively [3]. several researches has been done on TSE for different purposes including Single or multi-source traffic data analysis [4], Solving the shortest route (travel time) in the urban network [5, 6], Data noise removal and map matching in FCD's analysis [7, 8, 9], traffic forecast in the short or long future time period [10]. There are various techniques and methods for analysis, estimation and prediction of traffic characteristics such as kalman filter, HMM, FCE or learning method like spiking neural network [10,11,12].

Many researches have been done with sufficient amount of data for TSE and therefore they might not be effective on the missing data scope.

An algorithm is presented by [13] which uses the swarm optimization method with FCD to calculate the traffic parameter. Then it utilizes the Fuzzy Division to predict road traffic. Also, this paper achieved good performance for traffic congestion estimation and prediction, but it doesn't cover the missing states and requires further study on how to extend the method to cover missing states.

In [14] the volume of road urban traffic is estimated through multi-sourced data, using the couple matrix method and tensor factorization model to integrate GPS data and auxiliary data of tweeter. The model was accurate but high computational complexity makes it inefficient.

[15] presented a real-time algorithm for joining the parameters of traffics and noise by using the derivative-free multi-dimensional stochastic approximations of spall" method. This research offered a numerical computational model in the status of momentary changing the traffic conditions, for real time estimation. This is done by hybrid the Simultaneous perturbation gradient approximation method with the EM method to increase the speed of EM in the determination of traffic changes stage. Also, the Kalman filter method will be used to estimation of traffic parameters. The algorithm carries out traffic flow calculations for intersections, which makes it generally impossible to estimate road traffic correctly and can only be effective in managing intersection traffic.

A model is proposed for delivering the Multi-source data by using the weather Information alongside road traffic with an equation of linear ratio between weather features [16]. This study is limited to speed prediction during snowy weather.

A grid method to estimate and predict traffic via the density-based spatial clustering is presented by [17]. In this method, the road map is classified as a grid along the road. Moving on the road will be moving on the grid square which is done in discrete time. This allows the neighboring of the vehicle to be defined in the neighborhood of a two-dimensional grid, and they present an algorithm for Pattern Analysis by using the DBSCAN. The algorithm carries out traffic flow calculations for intersections, which makes it generally impossible to estimate road traffic correctly and can only be effective in managing intersection traffic.

Several studies have been done on Hidden Markov Model [1, 18, 19, 20, 21, 22, 23].

[18], detected traffic congestion and incidents from real-time GPS data collected from GPS trackers or drivers‟ smartphones. It introduced a new clustering

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

method without any prediction model. The model is limited to detection of congestion and incidents.

A study by [19] analyzed traffic through traffic prediction algorithm and considered vehicles that intend to stop on the road via using the speed data and location. This research used the Random Decision forest and Support Vector Machine. Road risk analysis and management, plus helping the advanced driving system are its significant benefits. However, the method is limited to predict the route and stopping intent of human-driven vehicles at urban intersections.

Jain, A. [20], represented real-time algorithm for traffic estimates and road conditions for car maneuvers by the Input-output HMM method which predicts road conditions three and a half seconds before the car's maneuvers. This method comes with 80% matching with real-time mode. The weakness of the method is that it is confined to anticipation of maneuvers.

[21] proposed a method at the microscopic level which utilized non-parametric kernel-based approach. The vehicle's path on the map is defined by the weighting function for any possible motion, and the kernel forms the weights of the movement path. This method is very similar to the method which is used the probability density function. Based on the probability weight formula, the main vehicle location will be characterized. Although the method is scalable and computationally efficient, but it can only estimate the travel time on a route and cannot estimate speed or traffic flow of the traffic.

A method introduced by [22] to estimate traffic signal phases from a sequence of maneuvers using the Bayesian learning algorithm. This article shows that the hybrid of Dirichlet distribution with learning will improve the performance of the algorithm. In this research, Learning is done by Bayesian algorithm. After the trained data is gained, the Viterbi algorithm is used to infer the traffic signal phases on the dataset. The method needs to be extended to traffic state estimation rather that traffic signal phase estimation which is disappearing because traffic signals have communication capabilities become instrumented with sensors.

A study by [24] focused on personal route prediction (PRP) problem. The PRP problem is presenting personalized path to person. This means that the most common algorithms finally suggest a route based on the shortest arrival time. This research is used the information of the previous car routes and first order Markov chain to generate car routes. The method must be extended to have other traffic variable such as traffic state.

The PSO method is proposed for urban traffic congestion by [25]. This study offered a TFP (traffic flow prediction). Which contain two section TVP (traffic volume prediction) and TSP (traffic speed prediction). Each section is calculated by the PSO (particle swarm optimization) method. In more precise terms, the main form of PSO, to estimate TSP and TVP, in the time (t+1) is written by the data in time (t). The prediction method is based on the two-river Fuzzy, which is both based on PSO. Moreover, it just determines the weight of two modes TVP and TSP and does not affect the learning of the algorithm, so traffic changes do not go well in this way.

A method based on Spiking Neural Network is presented by [26] to generate an algorithm to predict urban traffic for a long time. Prior to this article, major work has been done in the short term. This means that forecasts are made for the current time in places where traffic data is not available. However, this research has been conducted on the neural network for the long-term forecast (several days). In this method, two clustering is used, one in the daily mode where the traffic network is performed using a basic model “DBSCAN”. For long-term sequence of traffic, SNN method is used. The cluster proposed a prediction for future based on the sequence of observation in days gone by. In this study, the time to be prognosticated has been divided to the windows of time, $W = \{w_{m*x_0}, \dots, w_{m*x_m}\}$. The algorithm will be done in a window by using the SNN and proposed the prediction. Whenever the prediction exceeds an error within this range, the paper presents an easy algorithm that SNN clustering is broken and re-edited. So that the effects of things like crashes that are not considered are covered. It is worth noting that although clustering is provided as a preprocessing step, the only way to update the clusters is online, but in case of SNN error, the algorithm will have high complexity due to the updating clusters in every time.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The Bayesian filter which is used by [27], is one of the good methods in traffic analysis used the Bayesian filter (BF) model. To learn the BF filter method, this paper used the Gaussian process (GP) regression model. BF method uses the normal distribution to calculate the probability. To produce the best model for traffic chain by using corresponding sequence, the BF method is learned by GP method. One of the weaknesses of this method is due to the definition of the advanced model for the Kalman filter, which is the low speed algorithm that performs this algorithm in Macroscopic level. Moreover, the complexity of prediction of traffic by FCD data is high because of the matrix calculations at each stage of the Kalman filter.

A multiple model particle filter is used to accommodate the nonlinearity and switching dynamics of the traffic incident model, and the smoothing algorithm is applied to improve the accuracy of the estimate when data are sparse [28]. This method is only tested via synthetic data and, the California algorithm is not able to detect the incident when the inflow is 4000 veh/hour or below.

In [29], real time urban road traffic volume was estimated using "Coupled hidden Markov model". It used two chains, one contains GPS information and other chains produced by tweeter's data with the corresponding time in the coupled hidden Markov models. This work improved the speed of using the CHMM method by parallelization of the EM method. The algorithm is based on the real data of Chicago. They overcome the real time data sparsity by using the historical average traffic speed for each road link in the last 3 years. Twitter accounts about 33% of the crash report and report roughly 66% of traffic congestion. This makes the issue similar to GPS data with incidents. Nevertheless, due to the low diversity of Twitter data, its impact on the HMM will be low in the coupled method. This is considered to be the weakness of the article. Moreover, the social network data error is not measurable.

An HMM-based Traffic State Estimation model is proposed by [30] to estimate the missing traffic state using FCD. An algorithm based on clustering and pattern mining is proposed to find clusters with road segments that having similar traffic characteristics. A multi-clustering strategy is adopted to achieve a trade-off between clustering accuracy and coverage. The model estimated the missing traffic state with good accuracy. However, the model suffers from high computational cost due to time complexity of the embedded algorithm.

The tensor decomposition method has recently been presented and improved by [31]. In this study, a new method of tensor decomposition is presented to estimate the traffic flow pattern for the missing data. In this method, the tensor has been created like article [5]. But the tensor is decomposed by using r-NNCP method (regularization Non-negative CANE/PARAFAC factorization).

A tensor (Travel Time-Road Traffic Conditions) is decomposed using the Probabilistic traffic condition clustering by [5]. Furthermore, through using this method for each missing data in tensor, a cluster is presented which by the possible formulation of this cluster, the best possible replacement for the data is achieved.

In [34], the Extended Kalman filter is presented to forecast the current traffic condition. Moreover, the Kalman filter coefficients are updated to approximate the missing data with the best probability using fixed-camera data. However, this method is used for each step for road traffic density matrices of size o (N * M). The multiplication of the formula x (k + 1) = x ^ ~ (k + 1) + A (x (k) -x ^ ~ (k)) + Ww (k) and this process must be repeated in the loop until the expected error. Since it is written in the basic state, it performs a lot of calculations and, in the case of using FCD data instead of the fixed-camera data, it will not work because High FCD data size.

A method is presented by [38] to estimate the traffic speed which uses Schatten p-norm matrix completion algorithm to overcome the missing data issue. The weakness of the method is that the estimation error increases as the missing rate goes above 60 percent.

The missing data problem in the multi-source can be solved via using other data such as social network data [29], incident data [32] and the data of the constant sensor [33]. Furthermore, this paper just studies the TSE with the FCD which is considered as single source data. The importance is that whenever the estimation result of a method improves via using single source data, it will always be easy to hybridize the method by using auxiliary data. Moreover, the main challenge of the road Traffic State Estimation

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

is using only FCD data. Auxiliary data, despite solving the sparseness of data in the simple way, is not as reliable as FCD, and always the challenge of uncertainty apply to the problem which is ignored in this condition.

In the above researches, they tried to improve traffic estimation. Several researches studied the traffic for specific purposes such as incident prediction or congestion prediction which is limited to specific range. Some of them improve the traffic estimation with respect to missing data issue and mentioned the missing data rate [30,37]. The rest didn't highlight it in their work which shows that more researches must be conducted to enrich the traffic analysis and estimation considering the percentage of missing data. Besides all the effort in promoting traffic analysis, estimation and prediction, how to incorporate the traffic variables with other components remains as a challenging task. Furthermore, the methods in the related work suffer from high computational complexity which reduces the performance. In this paper, a new model is proposed to improve estimation accuracy without compromising the performance which use traffic flow as the variable and combine it with timeslot parameter to achieve better estimation of the missing states. Furthermore, the proposed model is tested with different missing rates which is discussed in the section 4.

3. PROPOSED MODEL

In this section, the proposed model is explained. There are some preprocessing steps which are applied before implementing the method. In Figure 1, the preprocessing steps and the proposed model are illustrated. In the preprocessing part, noise in the FCD will be eliminated by using the map matching algorithm which is proposed by [1]. Map matching is the process to match a sequence of real-world GPS coordinates into a digital map. As it is shown in Figure 1, the mapped data is extracted from the pre-processing step and is used to construct the traffic flow matrix.

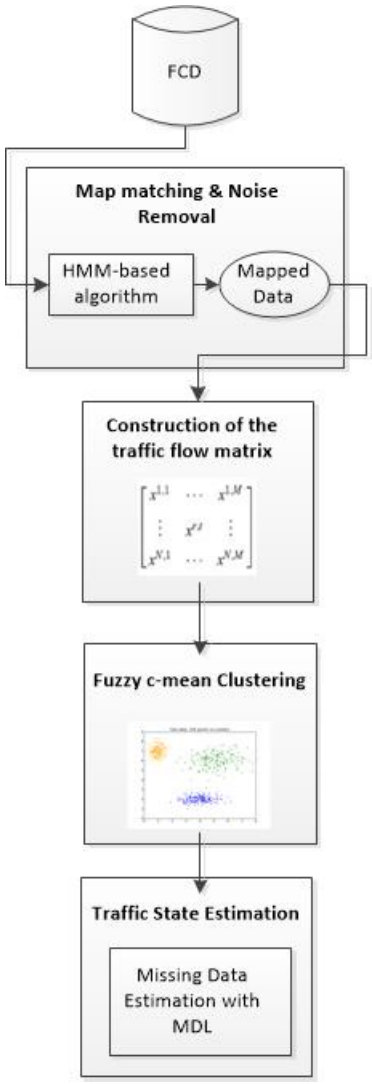


Figure 1: Fuzzy c-mean and MDL Model

The first step of proposed method is to construct the traffic flow matrix which is prerequisite for Fuzzy C-Mean clustering. Fuzzy C-Mean clustering was proposed by [38] and the steps will be explained in section 3.2. Fuzzy C-mean clusters the read segments based on similarity of their traffic flow. Eventually, final step estimates the missing traffic states by using MDL (Minimum Description Length). The overall flow of the model is shown in the figure 1.

In this research, the concept of existence of noise in the FCD is that if $\mathrm{T} = (P_1, P_2, .., P_n)$ is a GPS trajectory from a vehicle in the road network G, the error P_i in the original vehicle trajectory is considered as a GPS data error and will be removed. Consequently, lack of points in $\mathrm{T} = \{P_1, P_{2,}, \dots, P_m\}$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

in some segments of the vehicles' trajectory is considered as data sparseness.

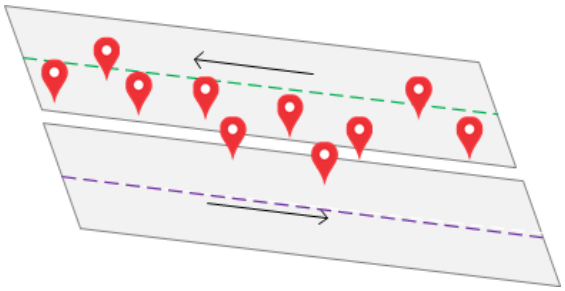


Figure 2: illustration of the GPS noise in the FCD

In Figure 2, the effect of noise in the data is shown. The red pins are the GPS points of a vehicles passing through the upper road. However, due to the errors some of GPS points are fallen on the other side of the road.

The approximation of the FCD data is the presentation of metric function to fitting a path from G to trajectory T which is earned from the vehicle's GPS data. If the measurement of this fitting is shown with f, then the form of this approximation is:

$$P = argmin_{p \in P_G} f(T, P) \quad (1)$$

Which P_G is path for the trajectory T. The process to choose f and solving the Equation (1) is mentioned as map matching algorithm [1].

The proposed model consists of three main parts including traffic flow matrix, Fuzzy C-Mean clustering and MDL to estimate the missing traffic states which will be explained in detail in the next section.

3.1 Construction of Road Traffic Flow Matrix

This section calculates the speed of traffic flow for all road segments, R = {r_n |n = 1,2, … , N} in the road network G at the traveling time t. this paper considers time as discrete in order to have easier statistical analysis and divides the time into 108 timeslots, T = {t_m |m = 1, 2, …, M}. The speed of traffic flow for segment r, at time slot t is shown as $x^{r,t}$, is equal to the average speed of all the vehicles passing the corresponding segment within each time slot.

$$X = \begin{bmatrix} x^{1,1} & \cdots & x^{1,t} \\ \vdots & x^{r,t} & \vdots \\ x^{N,1} & \cdots & x^{N,M} \end{bmatrix} \quad (2)$$

The traffic state for road segment r during all timeslots will be shown as X = {$x^{r,m}$ |m = 1,2, …, M}. As it is mentioned before, the FCD commonly suffer from lack of sufficient data which makes the computation of traffic state matrix difficult. The missing matrix elements needs to be estimated which in this paper it is referred to Traffic State Estimation. In the next section, the clustering step will be described.

3.2 Fuzzy c-mean clustering

In Figure 3, the clusters' center will be calculated for velocity of segments for each time slot. There are 108 time slots at 10-min intervals (e.g., the first time slot was 6:00–6:10 and the 12th time slot was 7:50–8:00). The number of clusters should reflect the expected number of road segment status i.e. 20 clusters represent with difference about 5 km/hours.

An extensively utilized objective function for Fuzzy C-means clustering is the weighted within-groups sum of squared errors J_m, which is employed for defining the restricted optimization issue as discussed in Equation 3:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \left\| x_i - c_j \right\| \quad (3)$$

where: 1≤ m ≤∞, m stands as a real number bigger than 1, u_{ij} refers to the extent of membership of x_i in the cluster j, x_i is the i^{th} component of d-dimensional recorded data, c_j is the center of the cluster, and ||*|| is any norm stating the resemblance between any measured data and the center. The partitioning of Fuzzy is performed via a repetitive optimization of the objective function, through the update of membership u_{ij} and the cluster centers c_j by Equations 4 and 5 respectively:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_j \right\|} \right)^{\frac{2}{m-1}}} \quad (4)$$

Journal of Theoretical and Applied Information Technology
31st July 2020. Vol.98. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m} \quad (5)$$

Input: A matrix of road segments' traffic flow speed at time t.

Output: K cluster of roads segments based on traffic flow speed.

1. Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$
2. At k-step: calculate the centers vectors $C^{(k)}=[c_i]$ with $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m}$$

3. Update $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_i - c_j\|}{\|x_i - c_j\|}\right)^{\frac{2}{m-1}}}$$

4. If $\| U^{(k+1)} - U^{(k)}\| < \varepsilon$ then STOP; otherwise return to step 2.

Figure 3: The algorithm to converge a local minimum of J_m.

When the $\max_{ij} \{| U_{ij}^{(k+1)} – U_{ij}^{(k)}| < \varepsilon\}$ the iteration will stop. K is the iteration step and ε is a termination criterion between 0 and 1.

3.3 Traffic flow state estimation

In Figure 3, the missing traffic flow states are estimated. The algorithm estimates the traffic flow states based on the similarity of other roads segments traffic flow which share the same clusters created by the Fuzzy C-mean. In order to find the missing traffic flow state belongs to which cluster at time t, MDL principle (Equation 6) is utilized to measure the degree of similarity between all road segments exist in the same cluster with missing traffic states at previous timeslots.

$$MDL = L(H) + L(D|H) \quad (6)$$

The main concept of MDL is that the best hypothesis (H) for a given Dataset (D) is the hypothesis which lead to the best compression of the data [36]. The MDL concept is used in this paper to estimate the best value for the missing traffic state. in this paper, the hypothesis (H) is estimate the missing traffic states by using the membership function u_{ij}. Then, all the road segments which have missing traffic states at any of the timeslots will be considered as Dataset (D). For each road segment which traffic state is missing, the MDL search the FCM clusters at previous timeslots in order to find the others road segments that share the same cluster with that road segment.

At this point, it is important to consider a minimum support for the number of previous timeslots which need to be checked in order to estimate the missing traffic states. It is suggested that the total number of timeslots divide by the total number of clusters can be used as the basis to choose the minimum number of supporting timeslots.

$$MinSup(T, CS) = T/CS \quad (7)$$

Finally, the cluster with maximum MDL is selected to represent the missing road traffic state.

Figure 4 shows the Traffic State Estimation algorithm.

Input: K clusters at time t_n.

Output: estimated missing traffic state at time t_n.

Road segment r need to estimate its traffic state at time T, previous n interval of times $T = \{t_1, t_2, t_3, \ldots t_n\}$, traffic flow state clusters set for previous n interval $CS = \{cs_1, cs_2, cs_3, \ldots cs_n\}$, each $cs_t = \{c_1, c_2, c_3, \ldots c_m\}$ where m number of clusters.

1. $\forall t_i , t_i \in T$
2. $\forall c_j , c_j \in CS^t$
3. $MDL_{ij} = \sum U_{ij}^t, \widehat{r^t}, r_j^t \in c_j^t$
4. Select cluster with max MDL to represent missing state $\widehat{x_r}$.

Figure 4: Traffic State Estimation Algorithm

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A simple example is shown in the Figure 5 to convey the concept of traffic flow distribution over time by using Fuzzy C-mean clustering and MDL.

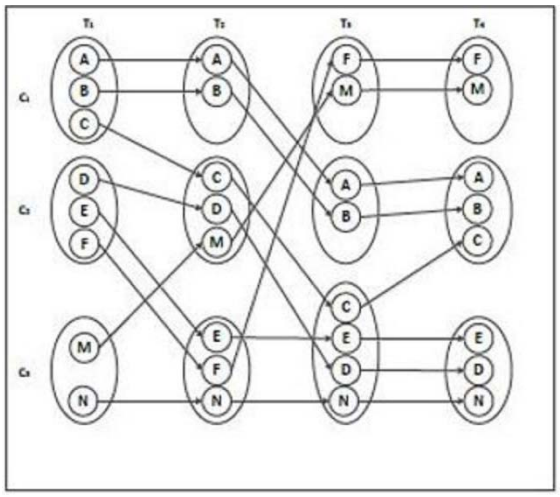


Figure 5: Traffic flow distribution over time by using Fuzzy C-mean clustering and MDL

Based on Figure 5, in order to calculate the membership for segment A at T_4, MDL is used to find the best cluster. At time T_1 segment A, B and C are in cluster C_1, at time T_2 segments A and B are in cluster C_1 again, at time T_3 segment A and B are in cluster C_2. As it can be seen in the Figure 5, segment B is the most similar segment to A which at cluster C_2 at T_4 with 3 degree of support. Whereas, the support degree for segment C is one. Consequently, among the 3 clusters at T4, C_2 would be the best candidate cluster for segment A.

4. EXPERIMENTAL RESULT

In this section, the findings are presented and the performances of the proposed method will be discussed. The proposed model was experimented using in MATLAB R2018a with Core i5-3470 3.20 GHz CPU and 16 GB RAM.

4.1 Dataset

The taxi trajectory data in Beijing during November 2012 served as the FCD data, obtained from 12,600 taxis.

The features which were used for experiment include, latitude, longitude, date and time, direction (heading).

4.2 Parameter Tuning

Based on Equation 7, the minimum number of supporting timeslots should be 6. Choosing the number of supporting timeslots below 6 may decrease the estimation accuracy. The grid search method has been used to find the best value for timeslots support. The experiment is tested for each missing rate with different timeslot support value. The timeslot support value for 0.93%and 0.463% of missing rate is set to 9 and for the rest of the missing rate percentages the value is set to 10.

Table 1 indicates the effect of timeslot support value on the MAE when the missing rate is 0.93%.

Table 1: The effect of timeslot support value on the MAE

Timeslot Support Value	MAE
1	4.3555
2	4.2995
3	4.2083
4	3.7905
5	3.7507
6	3.7383
7	3.7268
8	3.6535
9	3.5482
10	3.6363

The estimation error decreases as the number of timeslot support value increase.

Journal of Theoretical and Applied Information Technology
31st July 2020. Vol.98. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

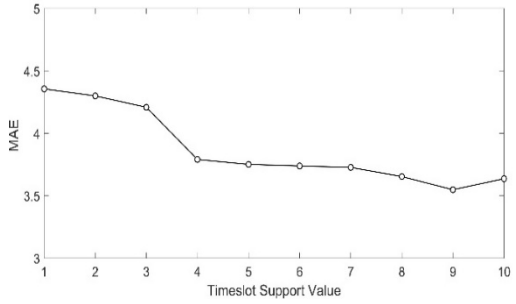


Figure 6: Estimation Error Corresponding To Different Timeslot Support Value

Figure 6 illustrates the decreasing gradient of MAE as the number of timeslot support value increases.

4.3 Performance Measurement

This work shows high accuracy to estimate the missing traffic flow state. Mean Absolute Error (MAE) is selected to measure the accuracy of our estimation model. As shown in table 1, MAE is reduced noticeably for different missing rate. The proposed model is compared with the HMM-based model in [30]. The calculation of MAE is shown in the Equation 8.

$$\text{MAE} = \frac{1}{N_{estim}} \sum_{1}^{N_{estim}} |\widehat{x_i} - x_i| \qquad (8)$$

Where $\widehat{x_i}, x_i$ are the estimated traffic state and real traffic state respectively, and N_{estim} is the total number of states estimated. The MAE decreases when the difference between the actual state and estimated state is reduced.

In Figure 7, the accuracy of the proposed model and HMM-based model are compared. Then, the proposed model estimated the missing traffic states of road segment using Fuzzy C-mean clustering (cluster based on similar traffic state) and MDL, while HMM-based model finds road segments with similar traffic states via clustering and frequent pattern mining. Our proposed model can estimate the missing traffic states better than the compared model for all missing rates.

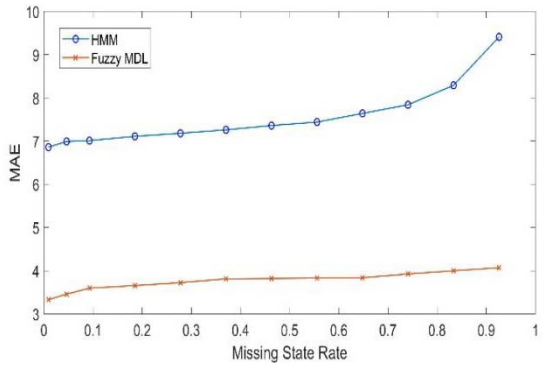


Figure 7: Comparison Of Accuracies Of The Fuzzy-MDL And HMM-Based Models

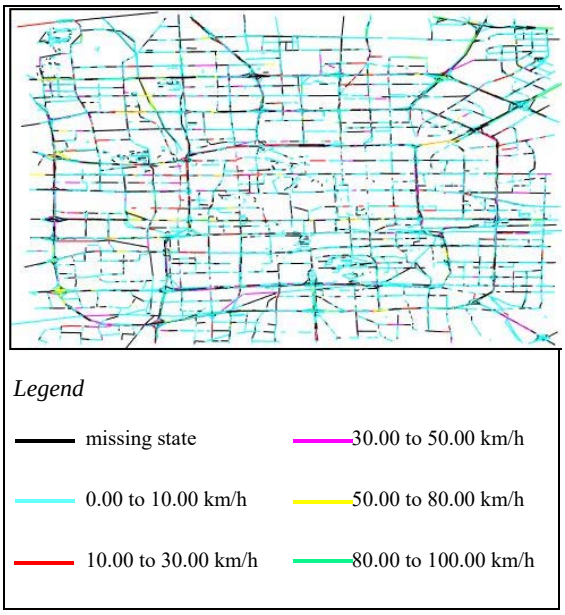
In Table 2, the experiment results of the proposed model are shown. The missing rates are from 0.93% to 92.59%. The amount of MAE for the proposed model and the HMM-Based model are 3.5482 and 6.86 respectively when the missing rate is 0.93. The implementation results demonstrate that the proposed model attained the lowest MAE in comparison to all missing rates.

Table 2: Traffic Flow State Estimation Accuracy Comparison

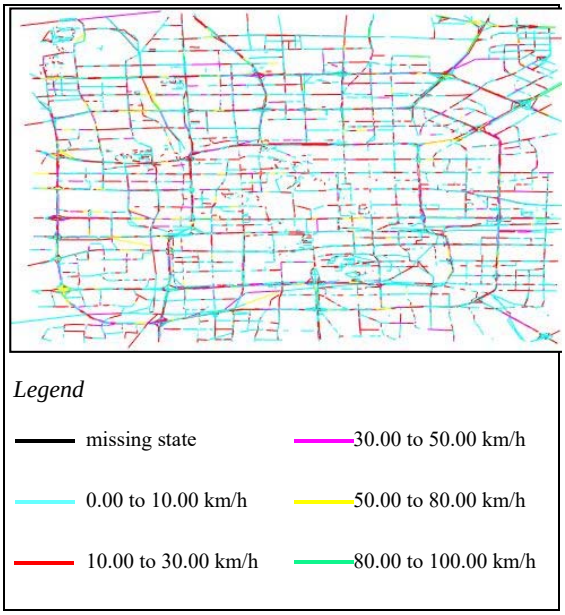
R_{miss} %	Timeslot support	MAE_{HMM} $(\frac{km}{h})$	$MAE_{Fuzzy-MDL}$ $(\frac{km}{h})$
0.93	9	6.86	3.5482
4.63	9	6.99	3.5587
9.26	10	7.01	3.5957
18.52	10	7.11	3.6558
27.78	10	7.18	3.7123
37.04	10	7.26	3.7575
46.3	10	7.36	3.778
55.56	10	7.44	3.8294
64.81	10	7.64	3.9208
74.07	10	7.84	3.9471
83.33	10	8.29	3.9988
92.59	10	9.41	4.065
97.22	10	11.96	4.127

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The results from Fuzzy c-mean achieve higher accuracy to cluster roads segments' traffic flow in comparison with HMM-based model which creates very big number of clusters based on graph theory and take a lot of time. Also, the estimation accuracy calculated based on MDL gives better results as this model is more suitable for trajectory rather than HMM-based model. In this paper, a promising model is proposed to estimate traffic flow state more accurately and efficiently.



(a) Before estimation



(b) After estimation

Figure 8: Traffic States Of Arterials In Beijing At The 50th Timeslot (2:20 Pm – 2:20 Pm) Before Estimation, And Figure 8(b) Shows The Traffic Condition Of The Arterials After Estimation.

Figure 8(a) illustrates the map of traffic states of arterials in Beijing at the 50th timeslot (2:20 pm – 2:20 pm) before estimation, and Figure 8(b) shows the traffic condition of the arterials after estimation.

The missing states are illustrated as black color in Figure 8(a) and after estimation the missing states are almost disappear in Figure 8(b). Table 2 shows the numerical expression of estimation error in form of MAE. As it can be seen in the table, the amount of MAE of the proposed model is reduced considerably compare to HMM-based model.

5. CONCLUSION

In this paper, a new model is proposed to estimate traffic flow state based on an efficient clustering algorithm using Fuzzy C-mean and MDL. Based on the experiment's results it is more suitable for Traffic State Estimation rather than HMM-based model. The aim of this study is to propose a model to improve estimation accuracy and reduce the estimation error which is achieved with good results. Based on the results illustrated in table 2, the model can estimate well for all the missing rates in comparison with the HMM-based model. The best results attained by the Fuzzy C-mean and MDL model using missing rates up to 83.33 is MAE<4 while the best result for HMM-based model using similar missing rates is MAE>6.86 which indicated the efficiency of the proposed model. Moreover, one of the advantages of the proposed model is that it utilizes a simple matrix computation and reduced the complexity of the algorithm. One of the directions for the future work is to come up with a formula for parameter tuning instead of using grid search method. Moreover, the effect of the proposed model on traffic state prediction can be considered as well. Also, proposing a new approach for computation using GPU as parallel processing technology instead of CPU is worth to research.

REFERENCES:

[1] Newson, P. and Krumm, J., 2009, November. Hidden Markov map matching through noise and sparseness. In Proceedings of the 17th ACM SIGSPATIAL international conference

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

on advances in geographic information systems (pp. 336-343). ACM.

[2] Seo, T., Bayen, A. M., Kusakabe, T., & Asakura, Y. (2017). Traffic state estimation on highway: A comprehensive survey. *Annual Reviews in Control*, *43*, 128-151.

[3] Guo, Q., Li, L. and Ban, X.J., 2019. Urban traffic signal control with connected and automated vehicles: A survey. Transportation research part C: emerging technologies.

[4] Wang, S., He, L., Stenneth, L., Philip, S.Y., Li, Z. and Huang, Z., 2016, June. Estimating urban traffic congestions with multi-sourced data. In 2016 17th IEEE International conference on mobile data management (MDM) (Vol. 1, pp. 82-91).

[5] Tang, K., Chen, S. and Liu, Z., 2018. Citywide spatial-temporal travel time estimation using big and sparse trajectories. IEEE Transactions on Intelligent Transportation Systems, (99), pp.1-12.

[6] Fusco, G., Colombaroni, C. and Isaenko, N., 2016. Short-term speed predictions exploiting big data on large urban road networks. Transportation Research Part C: Emerging Technologies, 73, pp.183-201.

[7] Li, Y., Huang, Q., Kerber, M., Zhang, L. and Guibas, L., 2013, November. Large-scale joint map matching of GPS traces. In Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 214-223). ACM.

[8] Mohamed, R., Aly, H. and Youssef, M., 2016. Accurate real-time map matching for challenging environments. IEEE Transactions on Intelligent Transportation Systems, 18(4), pp.847-857.

[9] Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G. and Huang, Y., 2010, November. T-drive: driving directions based on taxi trajectories.

In Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems (pp. 99-108). ACM.

[10] Laña, I., Lobo, J.L., Capecci, E., Del Ser, J. and Kasabov, N., 2019. Adaptive long-term traffic state estimation with evolving spiking neural networks. Transportation Research Part C: Emerging Technologies, 101, pp.126-144.

[11] Jin, J. and Ma, X., 2017. A non-parametric Bayesian framework for traffic-state estimation at signalized intersections.

[12] Sutarto, H.Y., Boel, R.K. and Joelianto, E., 2015. Parameter estimation for stochastic hybrid model applied to urban traffic flow estimation. IET Control Theory & Applications, 9(11), pp.1683-1691.

[13] Yang, Q., Wang, J., Song, X., Kong, X., Xu, Z. and Zhang, B., 2015, November. Urban traffic congestion prediction using floating car trajectory data. In International Conference on Algorithms and Architectures for Parallel Processing (pp. 18-30). Springer, Cham.

[14] Wang, S., He, L., Stenneth, L., Philip, S.Y., Li, Z. and Huang, Z., 2016, June. Estimating urban traffic congestions with multi-sourced data. In 2016 17th IEEE International conference on mobile data management (MDM) (Vol. 1, pp. 82-91).

[15] Zammit, L.C., Fabri, S.G. and Scerri, K., 2019. Real-time parametric modeling and estimation of urban traffic junctions. IEEE Transactions on Intelligent Transportation Systems.

[16] Tanimura, R., Hiromori, A., Umedu, T., Yamaguchi, H. and Higashino, T., 2015, September. Prediction of deceleration amount of vehicle speed in snowy urban roads using weather information and traffic data. In 2015 IEEE 18th International Conference on Intelligent Transportation Systems (pp. 2268-2273). IEEE.

[17] Liu, Y., Yan, X., Wang, Y., Yang, Z. and Wu, J., 2017. Grid Mapping for Spatial Pattern Analyses of Recurrent Urban Traffic Congestion Based on Taxi GPS Sensing Data. Sustainability, 9(4), p.533.

[18] D'Andrea, E. and Marcelloni, F., 2017. Detection of traffic congestion and incidents

from GPS trace analysis. Expert Systems with Applications, 73, pp.43-56.

[19] Gross, F., Jordan, J., Weninger, F., Klanner, F. and Schuller, B., 2016. Route and stopping intent prediction at intersections from car fleet data. IEEE Transactions on Intelligent Vehicles, 1(2), pp.177-186.

[20] Jain, A., Koppula, H.S., Raghavan, B., Soh, S. and Saxena, A., 2015. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In Proceedings of the IEEE International Conference on Computer Vision (pp. 3182-3190).

[21] Rahmani, M., Jenelius, E. and Koutsopoulos, H.N., 2015. Non-parametric estimation of route travel time distributions from low-frequency floating car data. Transportation Research Part C: Emerging Technologies, 58, pp.343-362.

[22] Gahrooei, M.R. and Work, D.B., 2015. Inferring traffic signal phases from turning movement counters using hidden Markov models. IEEE Transactions on Intelligent Transportation Systems, 16(1), pp.91-101.

[23] Ghahramani, Z. and Jordan, M.I., 1996. Factorial hidden Markov models. In Advances in Neural Information Processing Systems (pp. 472-478).

[24] Dai, Y., Ma, Y., Wang, Q., Murphey, Y.L., Qiu, S., Kristinsson, J., Meyer, J., Tseng, F. and Feldkamp, T., 2016, December. Dynamic prediction of drivers' personal routes through machine learning. In 2016 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1-8). IEEE. Model. Information, 8(3), p.95.

[25] Kong, X., Xu, Z., Shen, G., Wang, J., Yang, Q. and Zhang, B., 2016. Urban traffic congestion estimation and prediction based on floating car trajectory data. Future Generation Computer Systems, 61, pp.97-107.

[26] Laña, I., Lobo, J.L., Capecci, E., Del Ser, J. and Kasabov, N., 2019. Adaptive long-term traffic state estimation with evolving spiking neural networks. Transportation Research Part C: Emerging Technologies, 101, pp.126-144.

[27] Jin, J. and Ma, X., 2017. A non-parametric Bayesian framework for traffic-state estimation at signalized intersections.

[28] Wang, R., Work, D.B. and Sowers, R., 2016. Multiple model particle filter for traffic estimation and incident detection. IEEE Transactions on Intelligent Transportation Systems, 17(12), pp.3461-3470.

[29] Wang, S., Li, F., Stenneth, L. and Philip, S.Y., 2016, September. Enhancing traffic congestion estimation with social media by coupled hidden Markov model. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 247-264). Springer, Cham.

[30]. Wang, X., Peng, L., Chi, T., Li, M., Yao, X. and Shao, J., 2015. A hidden Markov model for urban-scale traffic estimation using floating car data. PloS one, 10(12), p.e0145348.

[31] Yang, S., Wu, J., Xu, Y. and Yang, T., 2019. Revealing heterogeneous spatiotemporal traffic flow patterns of urban road network via tensor decomposition-based clustering approach. Physica A: Statistical Mechanics and its Applications, 526, p.120688.

[32] D'Andrea, E. and Marcelloni, F., 2017. Detection of traffic congestion and incidents from GPS trace analysis. Expert Systems with Applications, 73, pp.43-56.

[33] Gahrooei, M.R. and Work, D.B., 2014. Inferring traffic signal phases from turning movement counters using hidden Markov models. IEEE Transactions on Intelligent Transportation Systems, 16(1), pp.91-101.

[34] Dhivyabharathi, B., Fulari, S., Amrutsamanvar, R., Vanajakshi, L., Subramanian, S.C. and Panda, M., 2015, September. Performance Comparison of Filtering Techniques for Real Time Traffic Density Estimation under Indian Urban Traffic Scenario. In 2015 IEEE 18th International Conference on Intelligent Transportation Systems (pp. 1442-1447). IEEE.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[35] Von Luxburg, U., 2007. A tutorial on spectral clustering. Statistics and computing, 17(4), pp.395-416.

[36] M. Riyadh, N. Mustapha, N. Sulaiman and N. Sharef, "ONF-TRS: On-line Noise Filtering Algorithm for Trajectory Segmentation Based on MDL Threshold", Journal of Artificial Intelligence, vol. 10, no. 1, pp. 42-48, 2016. Available: 10.3923/jai.2017.42.48.

[37] J. Yu, M. Stettler, P. Angeloudis, S. Hu and X. Chen, "Urban network-wide traffic speed estimation with massive ride-sourcing GPS traces", Transportation Research Part C: Emerging Technologies, vol. 112, pp. 136-152, 2020. Available: 10.1016/j.trc.2020.01.023.

[38] J. Bezdek, "Objective Function Clustering", Pattern Recognition with Fuzzy Objective Function Algorithms, pp. 43-93, 1981. Available: 10.1007/978-1-4757-0450-1_3.