IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 12, 2021, accepted April 8, 2021, date of publication April 26, 2021, date of current version June 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3075911

Predicting Traffic Flow Propagation Based on Congestion at Neighbouring Roads Using Hidden Markov Model

BAGUS PRIAMBODO[1,2], AZLINA AHMAD[1], AND RABIAH ABDUL KADIR[1]
[1]Institute of IR4.0, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia
[2]Faculty of Computer Science, Universitas Mercu Buana, Jakarta 11650, Indonesia

Corresponding authors: Bagus Priambodo (bagus.priambodo@mercubuana.ac.id) and Azlina Ahmad (azlinaivi@ukm.edu.my)

This work was supported in part by the Ministry of Education (Malaysia) through the Fundamental Research Grant Scheme under Project FRGS/1/2018/ICT02/UKM/02/8.

ABSTRACT Nowadays traffic congestion has become significantly worse. Not only has it led to economic losses, but also to environmental damages, wastage of time and energy, human stress and pollution. Generally, traffic congestion is a ripple effect of a road congestion on neighboring roads. When congestion occurs, it will propagate through the road network due to increasing traffic flow. One of the complexities of traffic congestion is unpredictability, thus it is difficult to represent traffic flows by numerical equations. One possible approach is to use the spatial historical data of traffic flow and relate them with traffic condition (congestion or clear) using statistical approach. Studies on traffic flow propagation generally involves visualization with real time GPS trajectory data to help analyze traffic flow propagation using human vision. Our research focuses on traffic flow pattern based on data from sensors without having information about the connected roads. We study spatial and temporal factors that influence traffic flow near a congested road in a neighboring area. Hence, our study investigates the relationship of roads in a neighboring area based on the similarity of traffic condition. Roads with high relationship with other neighbouring roads are identified by extracting spatial and temporal features using traffic state clustering. Grey level of co-occurrence matrix (GLCM) is utilized with spectral clustering to cluster road segments that have the same duration of road congestion in terms of day and time intervals. The emission probability is then calculated for prediction of traffic state impact of road congestion in neighboring area using Hidden Markov Model (HMM). We proposed HMM together with our clustering method to predict traffic state impact of road congestion. The experimental results show that the accuracy of prediction using the proposed HMM achieve 89%.

INDEX TERMS Intelligent transportation systems, Hidden Markov model, traffic flow propagation, road congestion prediction.

I. INTRODUCTION

Due to increase in population and number of private cars, traffic congestion has become considerably worse. Involving economic losses, environment damages [1], wastage of time and energy [2], also human stress and environmental pollution [3]–[5]. People need traffic guidance which can influence their driving behavior (including changing their driving habits and driving paths) [6]. Effect on one's driving can lead to changes in traffic flow state in the upstream and downstream of a road section and other road segments in the network. Two complexities of traffic congestion are dynamic and interrelated. In other words, traffic congestion will propagate from one road to neighboring roads [7], [8].

Most studies in traffic flow involved visualization to analysis of traffic congestion and traffic flow propagation. For example, a study was conducted to analyse spatiotemporal trajectory using visualization of real time trajectory data [9], [10], [11], [7], [12]. However, most studies did not predict impact of traffic conditions on neighboring roads. Usually, the prediction was used to manage, analyze the traffic and to prevent traffic congestions (minimizing their impact). There are several studies involved in predicting propagation of traffic congestion. Studies by [13]–[15], used

The associate editor coordinating the review of this manuscript and approving it for publication was Liang Hu.

graph approach to estimate congestion propagation. Another study by [16] used overlay polyline to detect congestion. All these studies used frequent subgraph to find frequent propagation tree and all of them used GPS trajectory data to identify travel route and connected roads. Other studies tried to predict propagation by clustering dynamically (re-clustering) based on time. The study which is related to our research is using graph approach by [17], [14], [15], and travel trajectory by [18]. However, building a connected graph and sub-connected graph recursively during a road congestion needs dataset that has information about connected roads. Since our dataset is taken from IoT sensors that has no information about connected road, we used another approach. We combined traffic flow similarity clustering and statistical approach to predict impact of congestion on traffic state in neighboring area. In our study, statistical Hidden Markov approach is used for predicting propagation of road congestion on other roads in the neighboring area. Many work has been done which applied HMM to predict traffic flow [2], [6], [19], [20], [21], [22]. A study which is similar to our work is by Jiang *et al.* [2]. They found that predicting traffic state based on high relationship roads is better than based on adjacency of roads. Their study used the average probability of traffic state in neighboring roads at certain time interval. For predicting the impact of road congestion, we extended from [2] by considering road congestion and time window as emission. Furthermore, we used similarity clustering of traffic state based on statistical feature from [19], [21]. We proposed traffic state clustering based on frequency of congestion occurrence at certain time interval. Grey level of co-occurrence matrix (GLCM) is utilized to cluster road segments that have same duration of road congestion (frequency of occurrence congestion) in terms of day and time intervals. The similarity of traffic state is then monitored to find the best k cluster using spectral clustering. The emission probability is then calculated for prediction of traffic state on neighboring roads using HMM. Our proposed clustering and model is then compared with HMM-Avg (average probability of neighboring roads at certain time intervals) based on multiple days clustering. The model is compared with predicting traffic state based on road congestion in neighboring area. There are two main contributions of this work. Firstly, we developed a traffic state clustering method for mining roads that have a high relationship with the target road. Secondly, we developed a Hidden Markov model for predicting traffic state impact of road congestion at certain time intervals.

The remainder of this paper is organized as follows. Section 2 describes the problem of predicting impact of road congestion. Section 3 discusses the construction of an HM-based traffic estimation model. It outlines the main steps of the proposed approach and presents an algorithm for real-time traffic estimation. Section 4 describes the implementation of the proposed model as well as a case study for assessing the accuracy of the model. Finally, Section 5 summarizes our findings and concludes the paper.

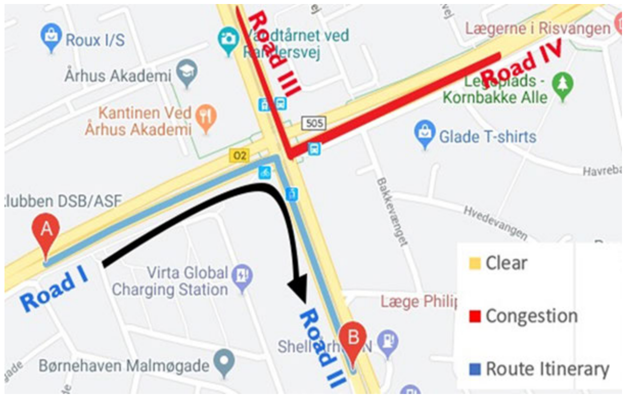

FIGURE 1. The situation of road congestion.

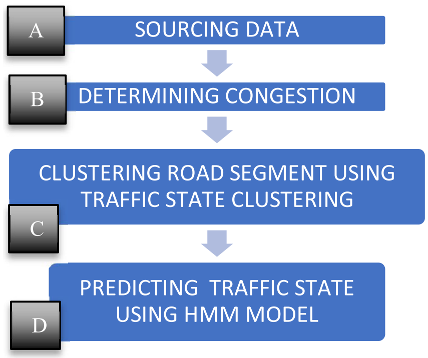

FIGURE 2. Methodology to predict traffic state impact of a road congestion.

II. PROBLEM DESCRIPTIONS

One of traffic management controllers' main responsibility is to predict state of traffic in short to medium range. The prediction can be used to manage traffic and to prevent traffic congestions and minimizing their impact. For example, travelling from point A to point B using road I and road II can be affected by traffic conditions at road III and road IV. The situation is depicted in detail in Fig. 1.

Based on Fig. 1, we can expect that the traffic flow on road III or road IV will also impact the traffic flow on surrounding roads like the road I and road II. The main objective of this study is to predict the traffic state of neighboring roads affected by a road congestion.

III. METHODOLOGY

Our methodology is explained in Fig. 2, descriptions for each step are explained in section 3.A - 3.D.

A. SOURCING DATA

In this study, we used the dataset obtained from the IoT traffic sensors located in Aarhus, Denmark [23]–[25]. The approximate number of sensors at this place is 449. However, in this paper we present only the results of six road locations with approximately 70 sensors located at this place as displayed in Fig. 3.

FIGURE 3. Map of six location of experiment in the city of aarhus, denmark.

TABLE 1. Details of data from sensor 158536.

Average Time	Average Speed	Time Median	Time	Vehicle Count
612	19	612	11:05:00	35
613	19	613	11:10:00	31
603	19	603	11:40:00	39
603	19	603	11:45:00	39

TABLE 2. Details of data from sensor 158536.

Time	Coordinates	Normal Driving Time
10/1/2014 1:45:00 AM	Latitude: 56.234 Longitude: 10.125 Postal Code: 8382	NDT: 109 (kmh) Distance between sensors: 3255 (m) Duration: 100 (s) Road type: Major Road
11/13/2014 10:40:00 AM	Latitude: 56.213 Longitude: 10.144 Postal Code: 8200	

For instance, the sensor at location A is identified by 158536. This sensor is placed at SÃ ftenvej Street, Aarhus city and at rhusvej Street, Hinnerup city, Denmark. The distance between both the sensors is 2061 metres. We conducted the experiment using average speed and time to calculate the congestion index. Example of traffic data taken from this sensor are presented in Table 1.

B. DETERMINING CONGESTION

Determination of road congestion levels is crucial [26]. As a result, various definitions of traffic congestion have been outlined. Traffic congestion rank was defined by Rothenberg [27] as the condition in which the number of vehicles on the road surpasses the carrying capacity of the standard road service level. Another study used congestion index by considering the saturation degree, travelling speed and a combination of both [28]. A different study accounted for the speed performance index by segmenting congestion level as four, three or two as needed [29]. In this study, the congestion index was determined for a given time interval to calculate similarity between roads to obtain the congestion level.

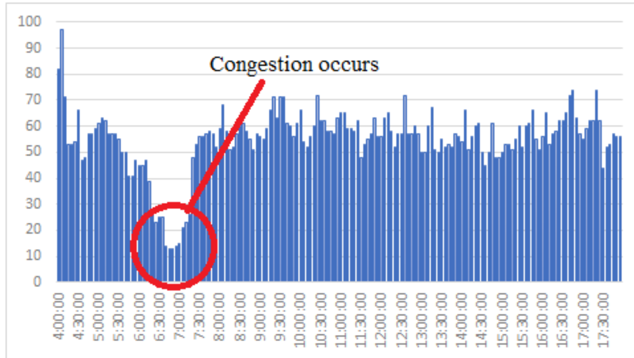

FIGURE 4. The average speed on road 158324 every five minutes.

Congestion index was calculated based on the travelling speed [28], with some adjustments. Instead of hourly calculation, congestion index was calculated every 20 minutes using Eq. 1. For example, for road 158324, congestion index was calculated from 04:00 to 10:00 AM as presented in Table 1I. For neighboring roads, congestion index was calculated every day from February 2014 until May 2014.

$$CI = \frac{NDT - Vavg_{int\_hours}}{NDT - Vmin_{int\_hours}} \times \frac{Volume_{int\_hours}}{Volume_{day}} \times 100 \quad (1)$$

NDT: normal driving time in kilometre per hour or speed limit, as shown in Table 2

$Vavg_{\,int\_hours}$: average speed in interval hours

$Vmin_{\,int\_hours}$: minimum speed in interval hours

$Volume_{\,int\_hours}$: number of vehicles in interval hours

$Volume_{\,day}$: number of vehicles in a day

As we can see from Fig. 4, road congestion occurred between 06:10 AM until 08:10 AM. From Table 2, it shows that the congestion index is between three (3.03) to five (5.41). In Denmark, the average speed of normal traffic in town is 50 km/hour [30]. Based on this information, we defined traffic congestion as the situation when average speed is below 50 km/h. As we can observed from Table 3, when the average speed is 50 km/hour, the congestion index value is around 3. Thus, for this study, we consider that a road is congested when the congestion index is above or equals 3.

C. TRAFFIC STATE CLUSTERING

Connected roads with both downstream and upstream are also used as factors in predicting traffic flow)[31], [32] [33] [34]. Other studies used var,ious method to determine spatial relationship between roads. Linear regression [35] [24], [25], and clustering method, k-Means [21], [38] and spectral clustering [2] have been widely used to group road segments. There is also a study using probabilistic Markov chain [39] to find similarity pattern between roads in spatial and time.

In this study, we use traffic state clustering to obtain the relationship between roads. Number of occurrences of traffic

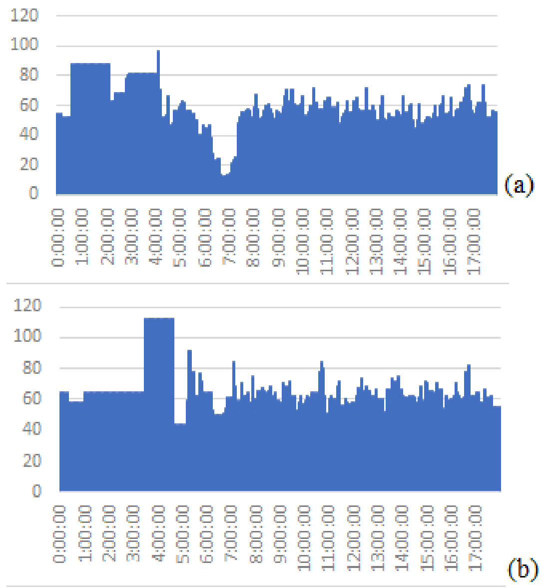

FIGURE 5. (a) Traffic pattern on weekdays, (b) Traffic pattern on weekends.

TABLE 3. Congestion Index (ci) on road 158324 from 05.40 – 09.20 am.

TIME	CI	AVG SPEED				VOLUME			
5:40-6:00	2.73	63	62	57	57	4	6	14	17
6:00-6:20	**3.03**	**57**	**55**	**50**	**50**	**12**	**12**	**15**	**13**
6:20-6:40	3.83	41	41	47	45	16	20	11	11
6:40-7:00	4.52	45	47	39	28	12	12	20	23
7:00-7:20	4.25	23	25	25	14	24	21	15	17
7:20-7:40	5.41	13	13	14	15	22	20	17	15
7:40-8:00	3.98	21	23	26	48	18	30	22	16
8:00-8:20	3.71	53	56	56	57	11	20	18	13
8:20-8:40	2.6	58	57	52	59	13	8	14	10
8:40-9:00	2.69	68	58	51	52	7	10	12	9
9:00-9:20	2.81	57	60	61	58	5	5	9	12

state for each road is calculated for clustering. There are two algorithms for feature extraction which can be used; Local Binary Pattern (LBP) and Grey Level Co-occurrence Matrix (GLCM). However, we selected GLCM since we want to cluster road segments based on traffic state (clear, congestion). If we use LBP, we have to divide the congestion index value into grayscale value (0-255), after which we need to compare neighboring value that is greater or less than the centre pixel value [40]. Furthermore, GLCM is a method for pattern recognition by studying the spatial correlation characteristics of grayscale. Since traffic flow pattern is formed by repeated occurrences of congestion distribution in spatial position, there exists a certain traffic state relationship between two matrix values at a certain distance in the traffic state space. This is the spatial correlation characteristics of traffic state. From Fig. 5, we can see that traffic congestion has similar patterns in terms of time and day intervals.

TABLE 4. Traffic congestion pattern in time and day on road 158536, road 173225, and road 158536.

Road 158536							
Date	**5:40**	**6:00**	**6:20**	**6:40**	**7:00**	**7:20**	**7:40**
1	0	0	1	1	1	1	1
2	0	1	1	1	1	1	1
3	0	1	1	1	1	1	1
Road 173225							
Date	**5:40**	**6:00**	**6:20**	**6:40**	**7:00**	**7:20**	**7:40**
1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1
Road 158386							
Date	**5:40**	**6:00**	**6:20**	**6:40**	**7:00**	**7:20**	**7:40**
1	1	1	1	1	1	1	1
2	0	1	1	1	1	1	1
3	1	1	1	1	1	0	1

Based on this pattern, we set our GLCM matrix with a horizontal offset of 4 to the right and vertical offset of 4 downwards. The offset is set to 4 because we want to find the traffic occurrence pattern on weekdays (Monday to Thursday). By referring to Table 4, 0 indicates clear state and 1 indicates congestion state.

Using GLCM, we obtained roads that have the same number of congestion occurrences in the cluster. There are some roads with the same number of occurrences but at different times (Refer Table 5). There are four roads (158415, 172602, 158865, 158954) that have the same duration of congestion with target road (158536) but occurred at different times. To overcome this issue, we added another time dimension in our GLCM clustering. We calculated the frequency of congestions at certain time intervals, and clustered using spectral clustering. The results are shown in Table 6. The clusters based on frequency are then intersected with clusters obtained using GLCM. The results of the intersection are our proposed traffic state clusterings as presented in Table 7.

The example of data for traffic congestion occurrences is given in Table 8, and the bar chart is shown in Fig. 6.

Our GLCM used contrast, homogeneity, correlation and energy. Contrast is a measure of intensity between a matrix value and its neighbour over the whole matrix [41] [42], [43]. Contrast is calculated using Eq. 2 and correlation of a matrix value to its neighbour is given by Eq. 3. Energy is the sum of squared elements in normalized GLCM as given by Eq. 4. Homogeneity is a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal. It is calculated using Eq. 5.

$$Contrast = \sum_{i,j} |i-j|^2 \, p(i,j)^2 \tag{2}$$

TABLE 5. The result obtained from GLCM spectral clustering on road 158536.

TIME	ROAD								
	158536	173225	158386	158415	172602	158983	158836	158865	158954
5:00	0	0		0	0	0	0	0	0
5:20	0	0	0	0	0	0	0	0	0
5:40	0	0	0	0	0	0	0	0	0
6:00	0	0	1	0	0	1	1	0	0
6:20	1	1	1	0	0	1	1	0	0
6:40	1	1	1	0	0	1	1	0	0
7:00	1	1	1	0	0	1	1	0	0
7:20	1	1	1	0	0	1	1	0	0
7:40	1	1	1	0	0	1	1	0	0
8:00	0	1	0	0	0	1	0	0	0
8:20	1	1	0	0	0	0	0	0	0
8:40	0	0	0	0	0	0	0	0	0
9:00	1	1	0	0	0	1	0	0	0
9:20	0	1	0	0	0	1	0	0	0
12:20	0	0	0	0	0	0	0	0	0
12:40	0	0	0	0	0	0	0	0	0
13:00	0	0	0	0	0	0	0	0	0
13:20	0	0	0	0	1	0	0	0	0
13:40	0	0	0	0	1	0	0	0	1
14:00	0	0	0	0	1	0	0	0	0
14:20	0	0	0	0	1	0	0	1	1
14:40	0	0	0	1	1	0	0	1	1
15:00	0	0	0	1	1	0	0	1	1
15:20	0	0	0	1	1	0	0	1	1
15:40	0	0	0	1	1	0	0	1	1
16:00	0	0	0	1	1	0	0	1	0
16:20	0	0	0	0	1	0	0	0	0
16:40	0	0	0	0	0	0	0	0	0

TABLE 6. The result obtained from frequent clustering on road 158536.

TIME	ROAD							
	158536	173225	158324	158386	158595	158983	158836	159014
5:00	0	0	0	0	0	0	0	0
5:20	0	0	0	0	0	0	0	0
5:40	0	0	0	0	0	0	0	0
6:00	0	0	0	1	0	1	1	0
6:20	1	1	0	1	0	1	1	1
6:40	1	1	1	1	1	1	1	0
7:00	1	1	1	1	1	1	1	1
7:20	1	1	0	1	0	1	1	1
7:40	1	1	1	1	1	1	1	1
8:00	0	1	0	0	1	1	0	0
8:20	1	1	0	0	0	0	0	0
8:40	0	0	0	0	0	0	0	0
9:00	1	1	0	0	0	1	0	0
9:20	0	1	0	0	0	1	0	0

TABLE 7. The result obtained from traffic state spectral clustering on road 158536.

TIME	ROAD				
	158536	173225	158386	158983	158836
5:20:00	0	0	0	0	0
5:40:00	0	0	0	0	0
6:00:00	0	0	1	1	1
6:20:00	1	1	1	1	1
6:40:00	1	1	1	1	1
7:00:00	1	1	1	1	1
7:20:00	1	1	1	1	1
7:40:00	1	1	1	1	1
8:00:00	0	1	0	1	0
8:20:00	1	1	0	0	0
8:40:00	0	0	0	0	0
9:00:00	1	1	0	1	0
9:20:00	0	1	0	1	0
9:40:00	0	0	0	0	0

TABLE 8. Number of occurrences of congestion.

Roads	5:00	5:20	5:40	6:00	6:20	6:40	7:00	7:20
158536	16	15	26	41	34	40	32	25
173225	19	18	35	41	42	39	39	34
158324	20	17	34	38	32	27	23	19
158386	22	23	31	44	41	35	29	23
158595	2	3	16	24	19	40	35	21
158715	12	19	22	31	37	40	28	22
158983	15	17	18	35	39	39	40	26
158836	19	14	25	45	45	41	33	24
159014	4	4	6	23	31	27	35	25

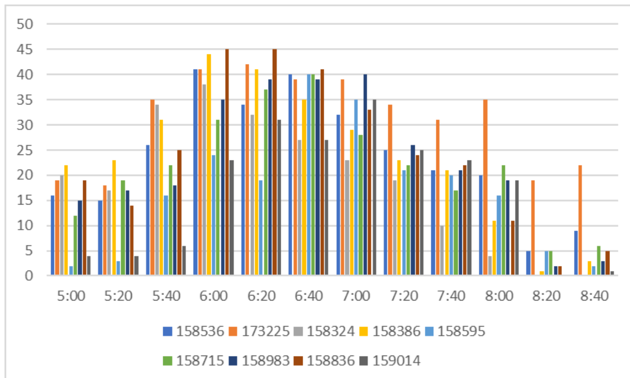

FIGURE 6. The frequency of congestions on road 158536 at different times.

$$Correlation = \frac{(i-\mu i)\,(j-\mu j)\,p(i,j)}{\sigma_i \sigma_j} \quad (3)$$

$$Energy = \sum\nolimits_{i,j} p(i,j)^2 \quad (4)$$

$$Homogeneity = \sum\nolimits_{i,j} \frac{(j-\mu j)\,p(i,j)}{1+|i-j|} \quad (5)$$

A spectral clustering algorithm is adopted to divide the road segments into clusters with similar traffic characteristics based on GLCM and frequency of congestions at certain time interval. The spectral clustering is used to obtain road segments that are located in the same cluster within the target road. Using this spectral clustering algorithm, the set of points in an arbitrary feature space can be represented as a complete weighted undirected graph $G\ (V, E)$. The vertices of the graph G are the points in the feature space and the weight W_{ij} of edge (v_i, v_j) in E is a measure of the similarity between vertex v_i, v_j. However, it is difficult to determine a suitable number of clusters for road segment clustering [2]. A study by [44]

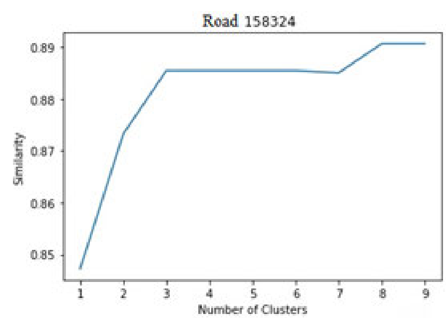

FIGURE 7. Similarity of roads between road 158324 and members of cluster.

used Davies–Bouldin (DB) index to find the best k cluster. Clustering scheme with the smallest DB value represents the optimal clustering performance. DB Index is the sum of square within clusters and sum of square between clusters.

TABLE 9. Result of prediction on neighboring roads 158324.

Traffic State Clustering			Multiple days spectral clustering	
Roads	Proposed HMM	HMM-Avg	Roads	HMM-Avg
158386	0.925	0.814	158386	0.825
158536	0.886	0.806	158536	0.818
173225	0.866	0.788	158715	0.828
158715	0.884	0.816	158983	0.812
158983	0.880	0.804		
158836	0.917	0.801		
Average	0.893	0.805		0.821

TABLE 10. Result of prediction on neighboring roads 158536.

Traffic state clustering			Multiple days spectral clustering	
Roads	Proposed HMM	HMM-Avg	Roads	HMM-Avg
173225	0.874	0.799	158324	0.895
158386	0.862	0.825	158386	0.802
158715	0.924	0.828	158715	0.805
158983	0.911	0.815	158983	0.792
158836	0.849	0.809	158836	0.790
			159014	0.792
Average	0.884	0.815		0.813

However, in our case we considered only the similarity within clusters, which is the similarity of target roads with members of the cluster. In this case, we used percentage of traffic state similarity of target road with members of the cluster to find the optimal number of k clusters. We expressed similarity between roads using Eq. 6, whereas similarity of target road and roads within cluster is expressed using Eq. 7 [45]. The similarity of traffic state is represented as line chart as shown in Fig. 7.

$$Similarity = \frac{1}{n}\sum\nolimits_{j=2}^{n}\left(Road_1 == Road_j\right) \quad (6)$$

$$Cluster\ Similarity = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\text{Similarity}_i\right) \quad (7)$$

Selecting a very high number for k may result in zero intersection between clusters. We selected the lowest number for k with the highest similarity and occurrence. For example, from Fig. 7, even though similarity occurred at 3, 4, 5, 6 and 7, however, we selected 3 as the value of k rather than 8, since it is the lowest.

D. PREDICTING TRAFFIC STATE USING HIDDEN MARKOV MODEL (HMM)

HMM has been widely used for prediction of traffic state. The observation state is used to predict the traffic state in a road segment. Previous studies used HMM for predicting traffic state with Variable Message Sign (VMS) as observed state [6]. Other studies by [19] [21] predict traffic state on freeway using HMM but using time window as observed state. Another study by [20] used HMM to predict traffic speed on highway whereas a study by [18] used HMM with double layer hidden state to predict vehicle trajectory. A study that is similar with our work is by [2]. They predict traffic state based on average speed on neighboring roads. The neighboring roads is obtained by performing multiple spectral clustering on different days. Observation state is calculated based on the average of emission probability in neighboring roads at certain time intervals. In this paper, we referred to their model as HMM-Avg (Average probability of neighboring roads at time intervals).

HMM is a statistical model used to describe the Markov process with hidden (unobserved) states (for more thorough description see [46], [47]). HMM can be presented as a form of dynamic Bayesian network.

There are three fundamental problems that are mainly solved using HMM [46]. They are:

The evaluation problem: Given the model $\lambda = (A, B, \pi)$ and observation sequence $O = O_1, O_2, \ldots, O_T$, determine the likelihood $P(O|\lambda)$.

The decoding problem: Given an observation sequence O and HMM model $\lambda = (A, B, \pi)$ discover the best hidden state sequence Q. Find the most likely state sequence in the model that produced based on observations.

The learning problem: Given the model $\lambda = (A, B, \pi)$ and observation sequence $O = O_1, O_2, \ldots, O_T$, how should we

TABLE 11. Result of prediction on neighboring roads 158715.

Traffic state clustering			Multiple days spectral clustering	
Roads	Proposed HMM	HMM-Avg	Roads	HMM-Avg
158983	0.921	0.804	158983	0.792
158836	0.840	0.801	159014	0.792
159014	0.822	0.804	158355	0.893
158536	0.912	0.806	192466	0.658
173225	0.866	0.788	158595	0.788
158324	0.836	0.797	158475	0.941
158386	0.847	0.814	158536	0.795
			192574	0.778
			158324	0.895
			158386	0.802
			192627	0.897
			193430	0.919
			193376	0.924
			195790	0.911
			194905	0.924
			195817	0.833
			193322	0.901
			184998	0.901
			182712	0.929
			210067	0.819
			179202	0.750
Average	0.864	0.802		0.850

adjust the model parameters (A, B, π) in order to maximize $P\{O|\lambda\}$?

A $= a_{11}, a_{ij}, \ldots a_{nn}$: a transition probability matrix

B: emission probabilities.

π = initial probability distribution

In our case, HMM is used as a decoder to find the most likely state sequence in the model which are produced based on given observations. There are two kinds of observation state in our study; observed state (the state of time and target road state) and hidden state (the state of traffic flow on neighboring roads). We determined the implicit parameters of the process from the observable parameters, and then use these parameters to do prediction. Detailed explanation is given below:

1. The hidden states $S = \{S_1, S_2\}$, are represented following the Markov chain. Our hidden state is divided into two: clear state and congestion state. Congestion state is the condition when congestion index value is above or equals 3 and clear state is the condition when congestion index is less than 3 as explain in section III (A). The transition probability matrix $A = [a_{11}, a_{12}, a_{21}, a_{22}]$ of hidden states represents the probability of transition between hidden states.
2. The n observed states $O = \{O_1, O_2, \ldots, O_n\}$, denote the number of observable states which is associated with the hidden states in the model. Observed states

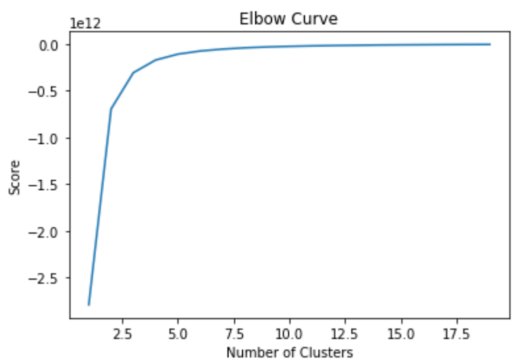

FIGURE 8. The elbow curve of time cluster in cluster of road 158324, cluster of road 158536, cluster of road 158715 and cluster of road 158954.

TABLE 12. Summary of the average accuracy of prediction on all location.

Roads	Traffic State Clustering Proposed HMM	Traffic State Clustering HMM-Avg	Multiple days spectral clustering HMM-Avg
158324	0.89	0.80	0.82
158536	0.88	0.82	0.81
158715	0.86	0.80	0.85
158954	0.88	0.88	0.87
192574	0.82	0.83	0.84
210013	0.82	0.83	0.85

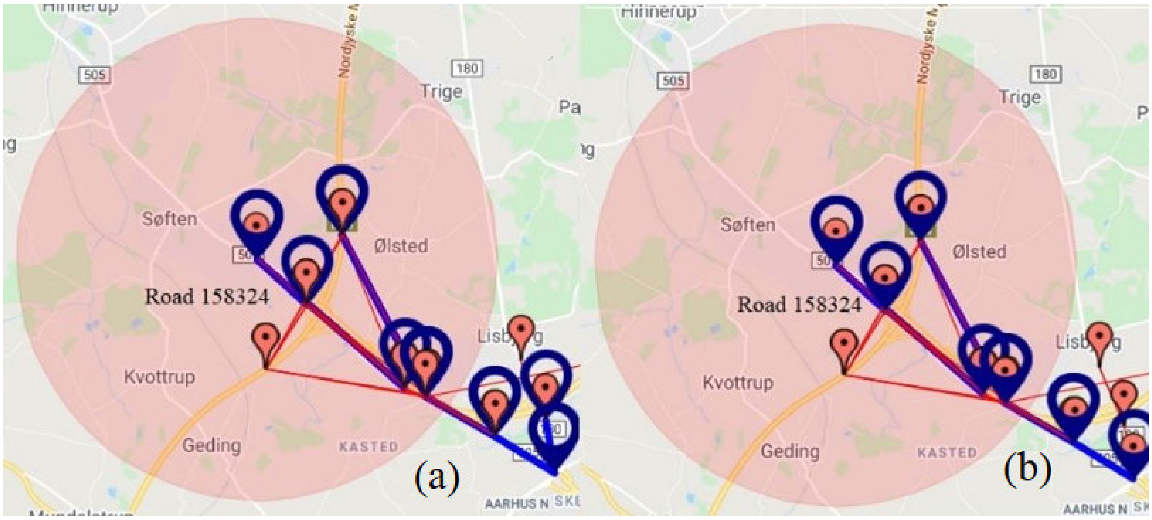

FIGURE 9. Relationship between roads for Road 158324 using (a) Traffic state clustering (b) Spectral with multiple days clustering.

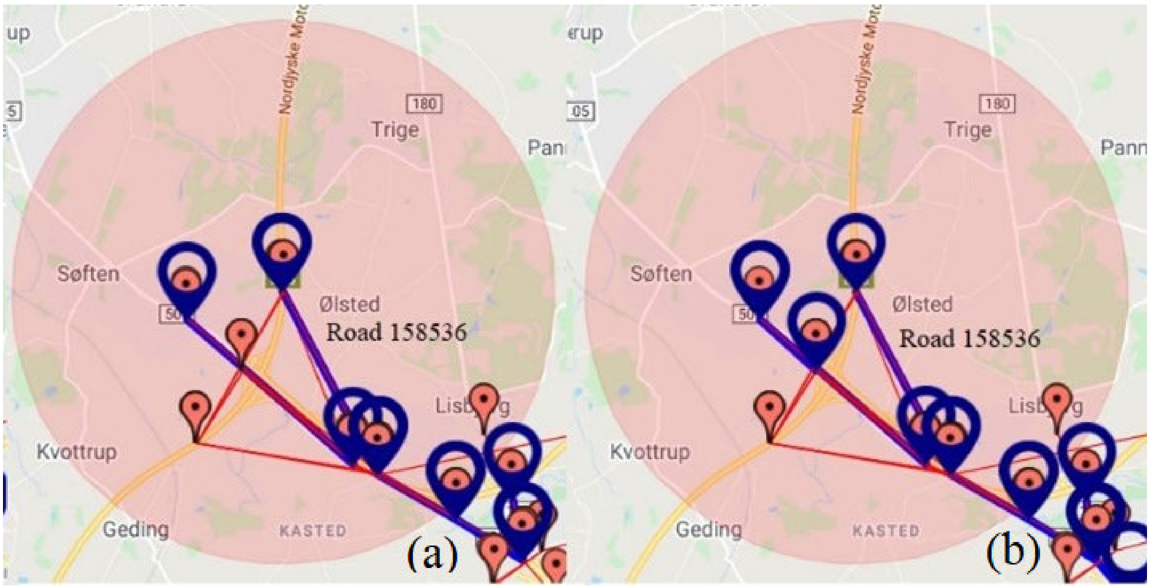

FIGURE 10. Relationship between roads for Road 158536 using (a) Traffic state clustering (b) Spectral with multiple days clustering.

describe the transition probability between hidden states and observed states in HMM. The transition probability between hidden states and observed states is called emission probability matrix B, where

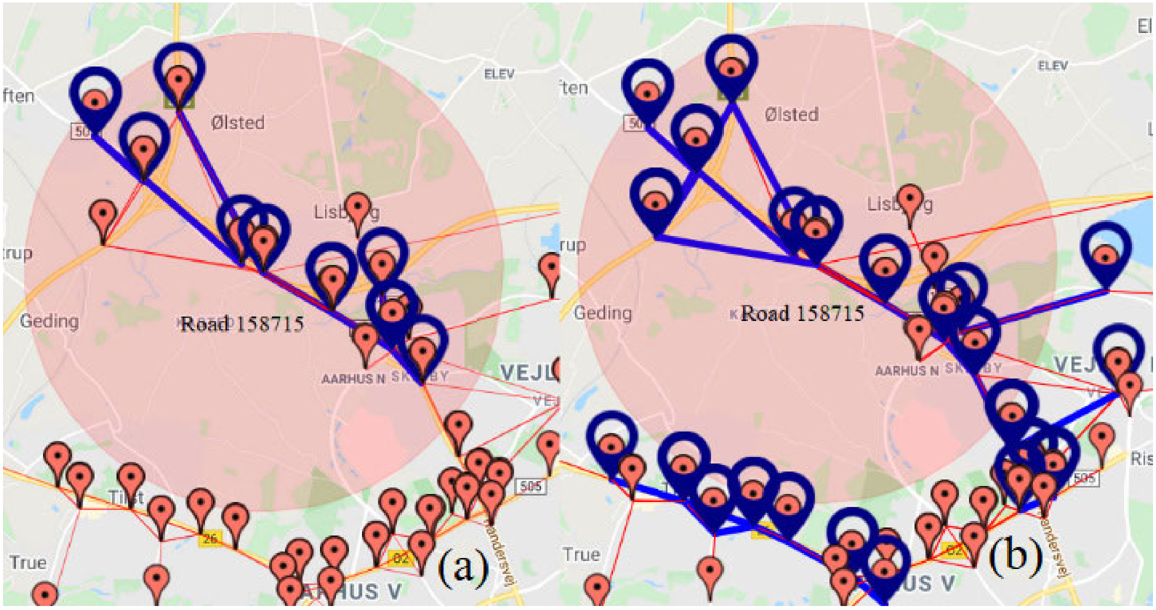

FIGURE 11. Relationship between roads for road 158715 using (a) Traffic state clustering (b) Spectral with multiple days clustering.

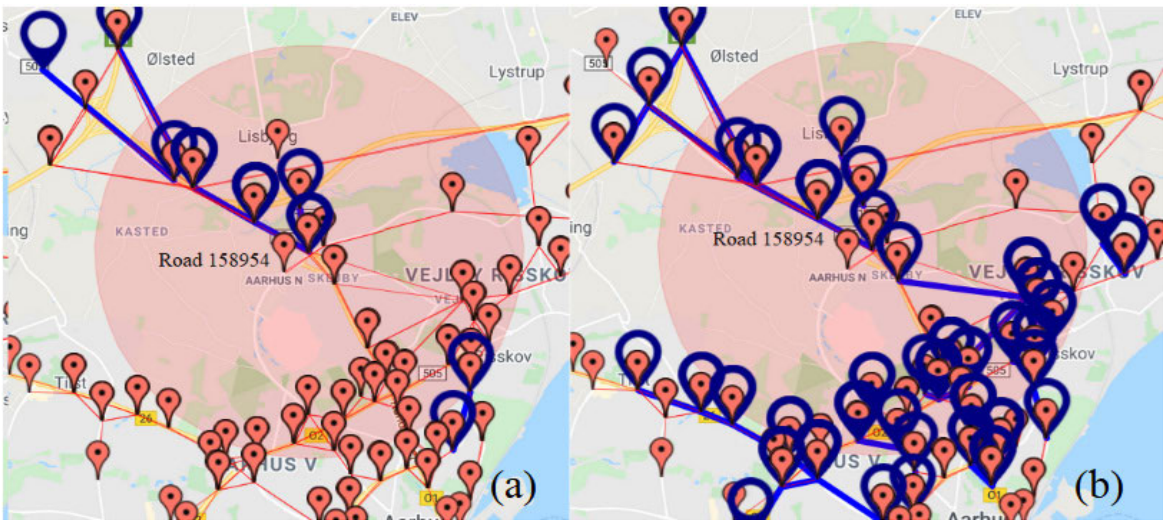

FIGURE 12. Relationship of roads for road 158954 using (a) Traffic state clustering (b) Spectral with multiple days clustering.

B $= b_{ij}P(O_i|\ S_j)$. It represents the probability of observed state O_i in the condition of hidden state S_j for $(1 \leq i \leq M,\ 1 \leq j \leq N)$ in timestamp t.

Our observation state is based on time cluster and traffic state of target road. After identifying the neighboring roads of the target road using traffic state clustering, we divide the traffic flow in the neighboring roads by time periods using K-Means. We determined the number k by observing the elbow curve of time cluster in four locations. Fig. 3 shows that the elbow point is 2.5 and become flat at five. However, the elbow point becomes really flat at six. Based on this information and the pattern of time cluster of average speed in a day [37], we divide time into six clusters. These six-time clusters are then used for defining our observation state. We divide each time cluster into two states; i) when target road has clear traffic flow, and ii) when target road is in a congestion state. From this, we obtained twelve observation states.

3. The initial state probability matrix π, denotes hidden state probability matrix at the time when initial timestamp t = 1. For example, for two hidden states, at time t = 1, $P(S_1) = \pi_1$ and $P(S_2) = \pi_2$, the initial state probability matrix $\pi = \pi_1, \pi_2$ and $S_{1,}S_2$ are hidden states.
4. The final step is the decoder. The dynamic Viterbi programming algorithm is then used for predicting the hidden traffic states sequences. Viterbi algorithm is used to find the most likely sequence of hidden states that can generate the given set of observations.

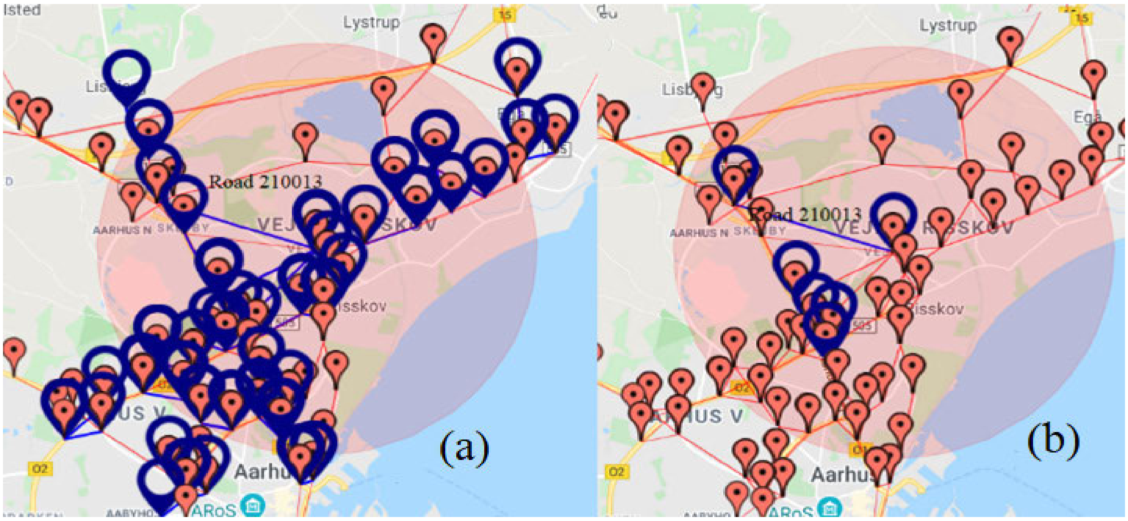

FIGURE 13. Relationship of roads for road 210013 using (a)Traffic state clustering (b) Spectral with multiple days clustering.

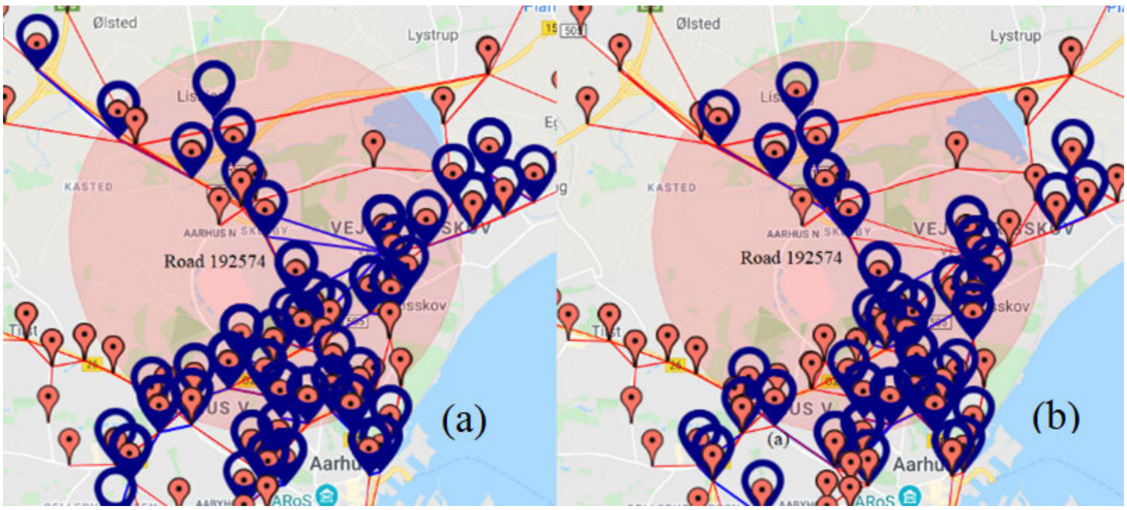

FIGURE 14. Relationship of roads for road 192574 using (a)Traffic state clustering (b) Spectral with multiple days clustering.

The state of target road and time cluster is treated as the observed sequence of events, whereas traffic state of road segment in a neighboring area is considered as the "hidden impact" of target road. The main task is to find the best traffic state sequence associated with a given state of target road and time. That is, predicting the optimal traffic state sequence impact of road congestion at a time interval. As defined by Rabiner [46], there could be more than one optimal criterion or there is no single optimal sequence. In our case, there are several paths through the hidden states (clear and congestion) that lead to the given sequence S. However, these paths do not have the same probability. The Viterbi algorithm is used to compute the most probable path. We implemented the Viterbi algorithm [48] to find the optimal traffic state sequence $S = \{S_1, S_2\}$ for a given observation sequence $O = \{O_1, O_2, \ldots, O_{12}\}$.

IV. RESULTS AND DISCUSSION

A. RESULTS

In this paper, we present the results of traffic state prediction at six (6) locations of roads namely road 158324, road 158536, road 158715. road 158954, road 210013 and road 192574. We define neighboring roads as roads that are located less than 5 kilometres from location.

We present our clustering results on maps to show the relationship between target road and its neighboring roads (Fig. 9, Fig. 10, Fig. 11, Fig. 12, Fig. 13, and Fig. 14). We also present the prediction results of our proposed HMM and HMM-Avg. Since we are predicting the state of traffic flow during weekdays, we computed the congestion index on weekdays starting from 15 Feb 2014 to 31 May 2014. Detailed explanation is given in section III (B). We split 20 percent of our dataset for testing the proposed HMM model. The prediction results of traffic state on neighboring

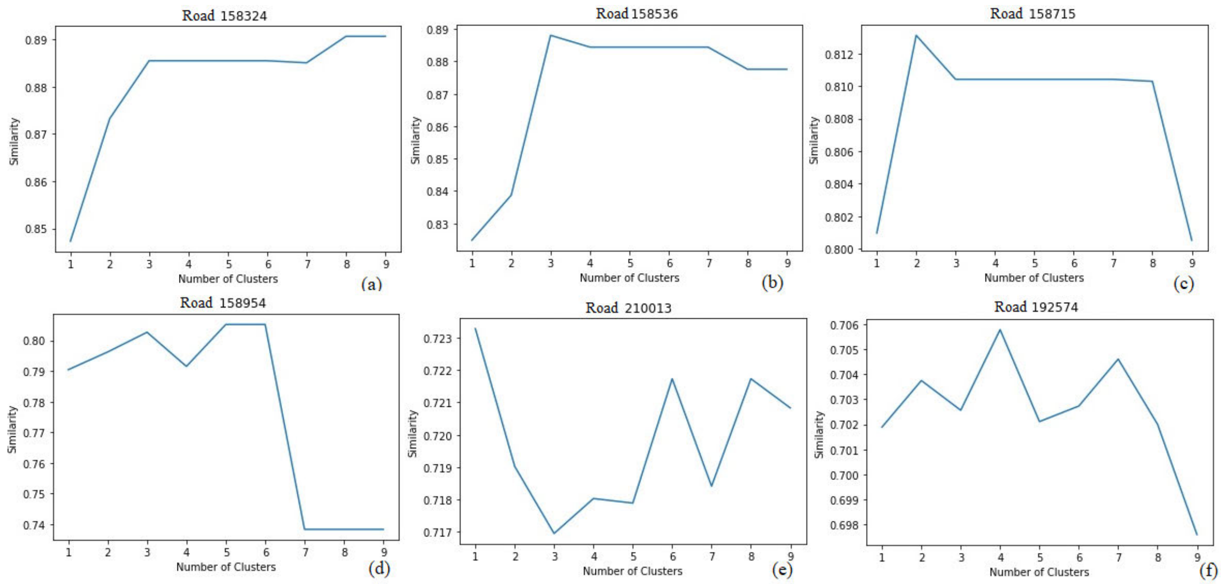

FIGURE 15. Similarity of roads based on number of clusters.

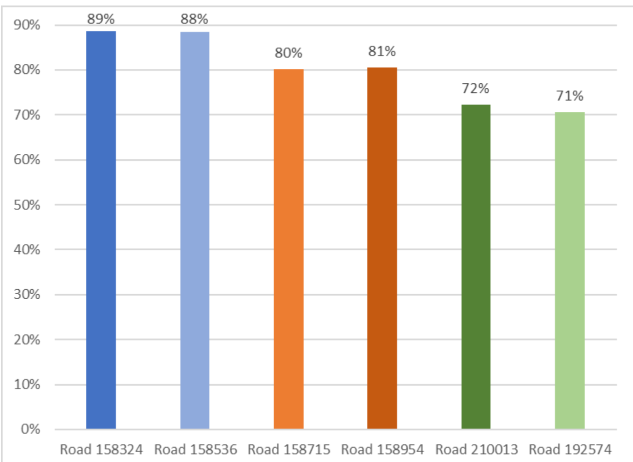

FIGURE 16. The similarity percentage of traffic state between target road and neighboring roads.

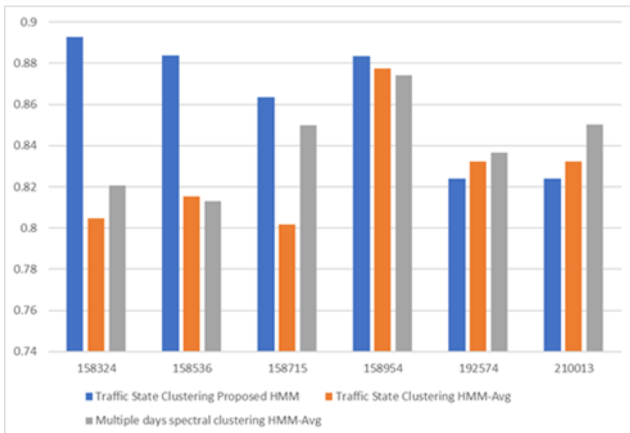

FIGURE 17. The comparison of traffic state prediction using proposed HMM and HMM-Avg.

roads are shown in Table 9 (road 158324), Table 10 (road 158536), Table 11 (road 158715), and the summary of all locations are presented in Table 12.

The relationship of roads are compared between our proposed traffic state clustering with spectral multiple day clustering and are then displayed on maps. Fig. 9, Fig. 10, Fig. 11, Fig. 12, Fig.13 and Fig.14 show neighboring roads of road 158324, road 158536, road 158715, road 158954, road 210013 and road 192574 respectively. The blue markers and blue lines are roads obtained from clustering which are considered as high relationship roads, whereas the red markers and the red lines are roads within radius of 5 km which have no relationship with target roads or excluded from the cluster. Figure 15 shows similarity of roads with member of clusters based on k. The explanation is given earlier in this paper using an example with Fig. 7.

B. DISCUSSION

From Fig. 9 and Fig. 10 for road 158324 road 158536 respectively, the results using traffic state clustering and multiple spectral cluster are similar. The roads are connected to each other and no road is separated from the target roads. From Fig.11 and Fig. 12, for road 158715 and road 158954 respectively, we observed that fewer roads are obtained when using traffic state clustering if compared with multiple days clustering. In other words, traffic state clustering filtered more roads than multiple days clustering. In Fig. 13 for road 210013, and Fig.14 for road 192574, different results were obtained. There are many roads separated or not connected with target roads. To investigate these differences, we summarized similarity of clusters from Fig. 15 to Fig. 16. From Fig.16, we verified the relationship between target road and neighboring roads on road 158324, road 158536, road 158715 and road 158954. The similarity of traffic state within these clusters is above 80%. However, for road 210013 and road 192574, the similarity of traffic state within these clusters is below 75%. These results explained why the traffic state clustering in road 210013 and road 192574 produced many roads that are separated from target roads.

Based on the experimental results, our proposed HMM gives better performance when compared with HMM-Avg. From Table 9, the highest prediction accuracy using our traffic state clustering method with our proposed HMM is 92.5% for road 158386. The highest average accuracy is 89.3% at neighboring roads of road 158324 (Table 9). Fig. 17 shows the comparison of accuracy between these two-clustering methods and two Hidden Markov models. Our proposed HMM produces superior result when a target road and neighboring roads have high relationship (similarity of traffic state within the cluster) as explain in Fig. 16. When traffic flow of target roads is less similar with neighboring roads, HMM-Avg shows better result.

V. CONCLUSION

The main objective of our study is to predict impact of a congested road on traffic condition in neighbouring roads. To achieve this, we study the relationship between a congested road with neighbouring roads. Relationship of roads with other neighbouring roads are obtained by extracting their spatial and temporal features using occurrence of congestions in terms of day and time. The extracted features are then clustered using spectral clustering. Neighbouring roads within the same cluster with target roads are then obtained as new set of neighbouring roads which is defined as high relationship roads. Our results show that our proposed HMM model with our proposed traffic state clustering performed better when compared with HMM-Avg. However, this method produces high results only when a congested road has high relationship with roads in a neighbouring area.

ACKNOWLEDGMENT

The authors would also like to extend the acknowledgement for the use of service and facilities of the Intelligent Visual Data Analytics Lab at IIR4.0, Universiti Kebangsaan Malaysia.

REFERENCES

[1] Y. Jiang, R. Kang, D. Li, S. Guo, and S. Havlin, "Spatio-temporal propagation of traffic jams in urban traffic networks," May 2017, *arXiv:1705.08269*. [Online]. Available: https://arxiv.org/abs/1705.08269

[2] X. Wang, L. Peng, T. Chi, M. Li, X. Yao, and J. Shao, "A hidden Markov model for urban-scale traffic estimation using floating car data," *PLoS ONE*, vol. 10, no. 12, pp. 1–20, 2015, doi: 10.1371/journal.pone.0145348.

[3] N. Petrovska and A. Stevanovic, "Traffic congestion analysis visualisation tool," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, Sep. 2015, pp. 1489–1494, doi: 10.1109/ITSC.2015.243.

[4] W. Hu, L. Yan, and H. Wang, "Traffic jams prediction method based on two-dimension cellular automata model," in *Proc. 17th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 2023–2028, doi: 10.1109/ITSC.2014.6958001.

[5] W. Zhou, W. Wang, X. Hua, and Y. Zhang, "Real-time traffic flow forecasting via a novel method combining periodic-trend decomposition," *Sustainability*, vol. 12, no. 15, pp. 1–22, 2020, doi: 10.3390/SU12155891.

[6] G. Zhu, K. Song, P. Zhang, and L. Wang, "A traffic flow state transition model for urban road network based on hidden Markov model," *Neurocomputing*, vol. 214, pp. 567–574, Nov. 2016, doi: 10.1016/j.neucom.2016.06.044.

[7] Z. Wang, M. Lu, X. Yuan, J. Zhang, and H. Van De Wetering, "Visual traffic jam analysis based on trajectory data," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2159–2168, Dec. 2013, doi: 10.1109/TVCG.2013.228.

[8] M. Saeedmanesh and N. Geroliminis, "Dynamic clustering and propagation of congestion in heterogeneously congested urban traffic networks," *Transp. Res. B, Methodol.*, vol. 105, pp. 193–211, Nov. 2017, doi: 10.1016/j.trb.2017.08.021.

[9] C. Tominski, H. Schumann, G. Andrienko, and N. Andrienko, "Stacking-based visualization of trajectory attribute data," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 12, pp. 2565–2574, Dec. 2012, doi: 10.1109/TVCG.2012.265.

[10] Z. Jianqin, Q. Peiyuan, D. Yingchao, D. Mingyi, and L. Feng, "A space-time visualization analysis method for taxi operation in Beijing," *J. Vis. Lang. Comput.*, vol. 31, pp. 1–8, Dec. 2015, doi: 10.1016/j.jvlc.2015.09.002.

[11] H. Bast, P. Brosi, and S. Storandt, "Real-time movement visualization of public transit data," in *Proc. 22nd ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2014, pp. 331–340, doi: 10.1145/2666310.2666404.

[12] X. Huang, Y. Zhao, C. Ma, J. Yang, X. Ye, and C. Zhang, "TrajGraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 160–169, Jan. 2016, doi: 10.1109/TVCG.2015.2467771.

[13] H. Xiong, A. Vahedian, X. Zhou, Y. Li, and J. Luo, "Predicting traffic congestion propagation patterns: A propagation graph approach," in *Proc. 11th ACM SIGSPATIAL Int. Workshop Comput. Transp. Sci.*, Nov. 2018, pp. 60–69, doi: 10.1145/3283207.3283213.

[14] Z. Chen, Y. Yang, L. Huang, E. Wang, and D. Li, "Discovering urban traffic congestion propagation patterns with taxi trajectory data," *IEEE Access*, vol. 6, pp. 69481–69491, 2018, doi: 10.1109/ACCESS.2018.2881039.

[15] H. Nguyen, W. Liu, and F. Chen, "Discovering congestion propagation patterns in spatio-temporal traffic data," *IEEE Trans. Big Data*, vol. 3, no. 2, pp. 169–180, Jun. 2017, doi: 10.1109/tbdata.2016.2587669.

[16] A. Keler, J. M. Krisp, and L. Ding, "Detecting traffic congestion propagation in urban environments—A case study with floating taxi data (FTD) in Shanghai," *J. Locat. Based Serv.*, vol. 11, no. 2, pp. 133–151, 2017, doi: 10.1080/17489725.2017.1420256.

[17] R. Besenczi, N. Bátfai, P. Jeszenszky, R. Major, F. Monori, and M. Ispány, "Large-scale simulation of traffic flow using Markov model," *PLoS ONE*, vol. 16, no. 2, 2021, Art. no. e0246062, doi: 10.1371/journal.pone.0246062.

[18] N. Ye, Y. Zhang, R. Wang, and R. Malekian, "Vehicle trajectory prediction based on hidden Markov model," *KSII Trans. Internet Inf.*, vol. 10, no. 7, pp. 3150–3170, 2016.

[19] Y. Qi and S. Ishak, "A hidden Markov model for short term prediction of traffic conditions on freeways," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 95–111, Jun. 2014, doi: 10.1016/j.trc.2014.02.007.

[20] L. Rapant, K. Slaninová, J. Martinoviè, and T. Martinoviè, "Traffic speed prediction using hidden Markov models for Czech Republic highways," in *Agent and Multi-Agent Systems: Technology and Applications* (Smart Innovation, Systems and Technologies), vol. 58, G. Jezic, Y. H. Chen-Burger, R. Howlett, and L. Jain, Eds. Cham, Switzerland: Springer, 2016, doi: 10.1007/978-3-319-39883-9_15.

[21] J. F. Zaki, A. Ali-Eldin, S. E. Hussein, S. F. Saraya, and F. F. Areed, "Traffic congestion prediction based on hidden Markov models and contrast measure," *Ain Shams Eng. J.*, vol. 11, no. 3, pp. 535–551, Sep. 2020, doi: 10.1016/j.asej.2019.10.006.

[22] Z. Li, C. Chen, Y. Min, J. He, and B. Yang, "Dynamic hidden Markov model for metropolitan traffic flow prediction," in *Proc. IEEE 92nd Veh. Technol. Conf. (VTC-Fall)*, Nov. 2020, pp. 1–5, doi: 10.1109/VTC2020-Fall49728.2020.9348555.

[23] R. Tonjes, P. Barnaghi, M. Ali, A. Mileo, M. Hauswirth, F. Ganz, S. Ganea, B. Kjargaard, D. Kuemper, S. Nechifor, D. Puiu, A. Sheth, V. Tsiatsis, and L. Vestergaard, "Real time IoT stream processing and large-scale data analytics for smart city applications," in *Proc. Poster Session, Eur. Conf. Netw. Commun.*, 2014.

[24] S. Kolozali, M. Bermudez-Edo, D. Puschmann, F. Ganz, and P. Barnaghi, "A knowledge-based approach for real-time IoT data stream annotation and processing," in *Proc. IEEE Int. Conf. Internet Things(iThings), IEEE Green Comput. Commun. (GreenCom) IEEE Cyber, Phys. Social Comput. (CPSCom)*, Sep. 2014, pp. 215–222, doi: 10.1109/iThings.2014.39.

[25] S. Bischof, A. Karapantelakis, A. Sheth, and A. Mileo, ''Semantic modelling of smart city data description of smart city data,'' in *Proc. W3C Workshop Web Things Enablers Services Open Web Devices*, 2014, pp. 1–5. [Online]. Available: http://www.w3.org/2014/02/wot/papers/karapantelakis.pdf.

[26] T. Afrin and N. Yodo, ''A probabilistic estimation of traffic congestion using Bayesian network,'' *Measurement*, vol. 174, Apr. 2021, Art. no. 109051, doi: 10.1016/j.measurement.2021.109051.

[27] Y. Zhang, N. Ye, R. Wang, and R. Malekian, ''A method for traffic congestion clustering judgment based on grey relational analysis,'' *ISPRS Int. J. Geo-Inf.*, vol. 5, no. 5, p. 71, May 2016, doi: 10.3390/ijgi5050071.

[28] W. Wang, R. Guo, and J. Yu, ''Research on road traffic congestion index based on comprehensive parameters: Taking Dalian city as an example,'' *Adv. Mech. Eng.*, vol. 10, no. 6, pp. 1–8, Jun. 2018, doi: 10.1177/1687814018781482.

[29] F. He, X. Yan, Y. Liu, and L. Ma, ''A traffic congestion assessment method for urban road networks based on speed performance index,'' *Procedia Eng.*, vol. 137, pp. 425–433, Jan. 2016, doi: 10.1016/j.proeng.2016.01.277.

[30] T. Hels, A. Lyckegaard, and N. Pilegaard, ''Evaluering af trafiksikkerhedstiltag—En vejledning,'' DTU Transp., Tech. Rep. 2011:1, 2011.

[31] C. Hu, K. Xie, G. Song, and T. Wu, ''Hybrid process neural network based on spatio-temporal similarities for short-term traffic flow prediction,'' in *Proc. 11th Int. IEEE Conf. Intell. Transp. Syst.*, Oct. 2008, pp. 253–258.

[32] B. Yu, X. Song, F. Guan, Z. Yang, and B. Yao, ''K-nearest neighbor model for multiple-time-step prediction of short-term traffic condition,'' *J. Transp. Eng.*, vol. 142, no. 6, 2016, Art. no. 04016018, doi: 10.1061/(ASCE)TE.1943-5436.0000816.

[33] H. Hong, W. Huang, X. Xing, X. Zhou, H. Lu, K. Bian, and K. Xie, ''Hybrid multi-metric K-nearest neighbor regression for traffic flow prediction,'' in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, Sep. 2015, pp. 2262–2267, doi: 10.1109/ITSC.2015.365.

[34] W. Min and L. Wynter, ''Real-time road traffic prediction with spatio-temporal correlations,'' *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 4, pp. 606–616, Aug. 2011, doi: 10.1016/j.trc.2010.10.002.

[35] Z. Liu, Z. Li, M. Li, W. Xing, and D. Lu, ''Mining road network correlation for traffic estimation via compressive sensing,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 1880–1893, Jul. 2016, doi: 10.1109/TITS.2016.2514519.

[36] B. Priambodo and A. Ahmad, ''Predicting traffic flow based on average speed of neighbouring road using multiple regression,'' in *Advances in Visual Informatics* (Lecture Notes in Computer Science), vol. 10645, H. B. Zaman *et al.*, Eds. Cham, Switzerland: Springer, 2017, doi: 10.1007/978-3-319-70010-6_29.

[37] B. Priambodo and A. Ahmad, ''Traffic flow prediction model based on neighbouring roads using neural network and multiple regression,'' *J. Inf. Commun. Technol.*, vol. 17, no. 4, pp. 513–535, Oct. 2018.

[38] F. F. A. J. F. Zaki, A. A. Eldin, E. S. Hussein, and F. S. Saranya, ''Framework for traffic congestion prediction,'' *J. Sci. Eng. Res.*, vol. 7, no. 5, pp. 1205–1210, 2016.

[39] J. Y. Ahn, E. Ko, and E. Kim, ''Predicting spatiotemporal traffic flow based on support vector regression and Bayesian classifier,'' in *Proc. IEEE 5th Int. Conf. Big Data Cloud Comput.*, Aug. 2015, pp. 125–130, doi: 10.1109/BDCloud.2015.64.

[40] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, ''Local binary patterns and its application to facial image analysis: A survey,'' *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 41, no. 6, pp. 765–781, Nov. 2011, doi: 10.1109/TSMCC.2011.2118750.

[41] A. Suresh and K. L. Shunmuganathan, ''Image texture classification using gray level co-occurrence matrix based statistical features,'' *Eur. J. Sci. Res.*, vol. 75, no. 4, pp. 591–597, 2012.

[42] I. Nurhaida, H. Wei, R. A. M. Zen, R. Manurung, and A. M. Arymurthy, ''Texture fusion for batik motif retrieval system,'' *Int. J. Electr. Comput. Eng.*, vol. 6, no. 6, pp. 3174–3187, 2016, doi: 10.11591/ijece.v6i6.12049.

[43] N. S. Fatonah, H. Tjandrasa, and C. Fatichah, ''Identification of acute lymphoblastic leukemia subtypes in touching cells based on enhanced edge detection,'' *Int. J. Intell. Eng. Syst.*, vol. 13, no. 4, pp. 204–215, Aug. 2020, doi: 10.22266/IJIES2020.0831.18.

[44] S. Yang, J. Wu, G. Qi, and K. Tian, ''Analysis of traffic state variation patterns for urban road network based on spectral clustering,'' *Adv. Mech. Eng.*, vol. 9, no. 9, pp. 1–11, 2017, doi: 10.1177/1687814017723790.

[45] B. Priambodo, A. Ahmad, and R. A. Kadir, ''Prediction of average speed based on relationships between neighbouring roads using K-NN and neural network,'' *Int. J. Online Biomed. Eng.*, vol. 16, no. 1, pp. 18–33, 2020, doi: 10.3991/ijoe.v16i01.11671.

[46] L. R. Rabiner, ''A tutorial on hidden Markov models and selected applications in speech recognition,'' *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989, doi: 10.1109/5.18626.

[47] Z. Ghahramani, ''An introduction to hidden Markov models and Bayesian networks,'' *Int. J. Pattern Recognit. Artif. Intell.*, vol. 15, no. 1, pp. 9–42, 2001, doi: 10.1142/S0218001401000836.

[48] G. Pulford, ''The Viterbi algorithm,'' *IET Seminar Dig.*, vol. 2006, no. 11359, pp. 53–65, 2006, doi: 10.1049/ic:20060556.

BAGUS PRIAMBODO is currently pursuing the Ph.D. degree with the Institute of IR4.0 (IIR4.0). He is also a Senior Lecturer with the Department of Information Systems, Faculty of Computer Science, Universitas Mercu Buana, Jakarta, Indonesia. His research interests include machine learning, computer vision, and big data analytics.

AZLINA AHMAD received the Ph.D. degree in multimedia from the University of Malaya. She is currently an Associate Professor with the Institute of IR4.0 (IIR4.0), Universiti Kebangsaan Malaysia (UKM), where she is also one of the Senior Research Fellows. She has been the principal investigator on research projects sponsored by the Ministry of Science and Technology and the Ministry of Education. She has experience in Project Management and is a certified Appraisal Team Members (ATM) for Capability Maturity Model Integration (CMMI) Level 2 from Carnegie-Melon University, USA. She has collaborated actively with researchers in several other disciplines, such as Islamic banking and finance, social science and humanities, and engineering. Her research interests include visualization, augmented and virtual reality, educational technology, and mathematics.

RABIAH ABDUL KADIR received the Ph.D. degree in computer science from the Universiti Kebangsaan Malaysia (UKM), in 2007. From September 2012 to August 2013, she was an Assistant Professor with Najran University, Saudi Arabia. Since June 2014, she has been a Research Fellow with the Institute of Visual Informatics, UKM. She is currently an Associate Professor and a Senior Research Fellow with the Institute of IR4.0 (IIR4.0), UKM. Her research interests include intelligent computing and computational linguistics. She is also active in research on big data analytics, semantic knowledge repository and extraction, and intelligent learning management systems.