Traffic Congestion Analysis Using Highway O-D Tollgate Data

Shi Fang1,*, Kaigui Bian1, Kunqing Xie1 and Xiaowei Ma1

1 School of Electronics Engineering and Computer Science
Peking University
Beijing, China

Abstract — Traffic congestion analysis is always of great interest for administrators of highway networks, as well as for the drivers. In this paper, we introduce a new approach to evaluate the congestion of each highway road section based on real-world tollgate O-D (Origin to Destination) data in Beijing highway network of China as an example. We propose the concept of the congestion index that measures the traffic intensity on both the section and road-level, which can be applied directly in different toll highway networks. A case study on the daily congestion pattern for Beijing highway roads was carried out, and a number of traffic congestion patterns are identified in this study.

Keywords - Traffic congestion index; traffic congestion pattern, real historical traffic data; toll highway

I. INTRODUCTION

The highway traffic congestion problem has garnered an increased attention from the public in recent years for its significant impact on environment and economics. Defining the “congestion (measurement) index” is proved to be imperative for describing and analyzing the highway congestion status.

Many features have been taken into the consideration of the congestion measurement. Most studies have agreed that a congestion measure should be accessed using following baselines:

(1) Easily understood and interpreted among different users.
(2) Precisely represents the quality of service and congestion magnitude.
(3) Capable to compare congestion levels across different metropolitan areas.

However, only a limited number of indices have been developed so far for standard measurements, which fail to offer the analytical capability to compare congestion levels across different metropolitan areas.

Some measurement indices are defined earlier for evaluating the traffic congestion on a particular road segment. The main indices include the daily vehicle-miles traveled per lane-mile (DVMT) [1], the traffic density and the traffic volume to rated capacity ratio (which is the most commonly used). The limitation of traffic volume to capacity ratio is that it cannot distinguish different congestion conditions, when the traffic demand is higher than the capacity. As an alternative, another measurement index called capacity adequacy index [2] is developed, which is defined as the highway rated volume capacity to the volume during peak hour ratio. Such an evaluation method is not applicable when deployed against an entire highway network. The widely adopted highway measurement method is the level of service (LOS) introduced by HCM [3], where the LOS classifies the highway traffic service into different quality levels based on the traffic volume to capacity ratio.

Later on, a few other indices are developed for analyzing the congestion in relatively large road networks. For instance, the roadway congestion index (RCI) is introduced to evaluate the traffic congestion in the entire urban areas, instead of a local street/block [4]. RCI combines the DVMT for freeway and arterial street system; meanwhile it allows the comparison between different traffic systems.

Another type of congestion measurement indices are built on basis of the travel time related factors. These approaches suggest that the travel time be sufficiently reliable to serve as the key index for congestion measurement from the perspective of a traveler or an operator. Such indices include: the travel time index (TTL) [5] - the peak hour travel time divided by the free flow travel time, and the planning time index [6] - the ratio of the 95th percentile peak period travel time to the free flow travel time. The assumption for this type of approaches is the knowledge of the precise information on individual travel time.

Most of the aforementioned measurement approaches require data obtained from additional equipment or hardware such as inductive-loop traffic detectors, video surveillance system or GPS-enabled floating-cars. Therefore, the performance of these approaches is highly dependent on, or restricted by the availability and coverage of these monitoring devices.

In a toll highway network, the tollgate data has been collected by the infrastructure provider and administrator without any additional monetary cost. Our research findings indicate that leveraging the tollgate O-D (origin-destination) data for the congestion analysis has the following advantages:

(1) Ease of calculating the travel time: Since the traffic is controlled by each tollgate and traffic demands can be closely monitored, one of the key index - the individual's travel time - can be easily obtained.
(2) Full coverage of the road network: The tollgate data have a nearly full coverage of the entire road network for a relatively long period. This enables us to examine the traffic behavior by different time-scale granularity (i.e. daily, weekly or yearly). Therefore we can identify the traffic patterns from these empirical data in the time-domain for congestion investigation.

In this paper, we present a new approach for traffic congestion analysis using the highway tollgate O-D (origin-destination) data. With such O-D information, we define a new congestion index to quantify the congestion level of each road network section. Since the traffic data is collected from a long-run operating transportation system, the purpose of this study is to identify the patterns for traffic congestion using the real-world O-D data, from both the section (segment) level and the road (group of sections) level. Based on the defined index, we carry out a case study in the Beijing arterial highway network to find the daily congestion pattern, using the K-means clustering.

The rest of paper is organized as follow: In Section 2 a brief classification is described for congestion analysis approaches. In Section 3 a new congestion index is defined for link-based and road-based measurement. In Section 4, we use a Beijing arterial highway network case study to retrieve the congestion pattern using K-means clustering. In Section 5 we conclude this research.

II. A TAXONOMY OF CONGESTION ANALYSIS METHODS

In this section, we present a taxonomy of existing approaches for congestion analysis, as shown in TABLE I. We broadly classify existing research into two categories of approaches, namely the traffic phenomenon and consequence approaches. The first category of approaches tries to determine the traffic congestion conditions using the instantaneous traffic parameters such as speed or density. However, the second approach puts more emphasis on the consequence caused by the congestion, such as delay, travel reliability. In this paper, we focus on the first category of approaches.

A. Traffic Density at the Road Section Level

The most important metrics to describe the traffic states are flow $f$, density $ρ$ and speed $v$. The flow is used to measure the number of vehicles that are passing a road cross-section at time $t$. The density $ρ$ can be calculated as following: $ρ = f / v$, which describes the number of vehicles per unit length of the road at instant time $t$. If we take a...
snapshot of the entire network at any time instance, we should be able to count how many vehicles on each road section. Therefore the amount of vehicles on a road section can be expressed as:

\[ q_{\text{section}}(t) = \sum q_{\text{section}}(\text{vehicle}_i, t) \]

(1)

Where \( q \) is a function whose returned value is equal to 1 only if the \( i \)th vehicle is traveling on the given road section at time \( t \); otherwise, the returned value is set to be 0.

\[ q_{\text{section}}(\text{vehicle}_i, t) = \{0, 1\} \]

(2)

1) Algorithm 1: Calculate Traffic Density Based On Average Travel Time Estimation Mode

The O-D data only contains information of the origin, destination tollgate, the entry and exit time. Additional knowledge is required to precisely identify whether the individual vehicle is driving on a particular section at a given time \( t \). Since we are capable to monitor all the vehicles O-D's information, we can estimate the value of \( q_{\text{section}}(\text{vehicle}_i, t) \) using the method below (the pseudo code is given in Algorithm 1 below).

Algorithm 1: Calculate of \( q_{\text{section}}(\text{Vehicle}_i, t) \).

<table>
<thead>
<tr>
<th>Require</th>
<th>The tollgate O, The tollgate D; The entering time for ( \text{Vehicle}_i ), The exiting time for ( \text{Vehicle}_i ); The section L, The instant time ( t ), Where ( \text{enterTime}(\text{Vehicle}_i) &lt; t &lt; \text{exitTime}(\text{Vehicle}_i) );</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensure</td>
<td>Is ( \text{Vehicle}<em>i ) traveling on section L, ( q</em>{\text{section}}(\text{Vehicle}_i, t) );</td>
</tr>
<tr>
<td>1: ( R = \text{FindPath}(O, D, \text{Vehicle}_i); )</td>
<td></td>
</tr>
<tr>
<td>2: if ( L \notin R ) then</td>
<td></td>
</tr>
<tr>
<td>3: ( \text{return } q_{\text{section}}(\text{Vehicle}_i, t) = 0; )</td>
<td></td>
</tr>
<tr>
<td>5: ( \text{speed} = \text{Distance}(R)/(\text{exitTime} - \text{enterTime}); )</td>
<td></td>
</tr>
<tr>
<td>6: ( \text{travelledDistance} = \text{speed} \times (t - \text{enterTime}); )</td>
<td></td>
</tr>
<tr>
<td>7: ( \text{distance} = 0; )</td>
<td></td>
</tr>
<tr>
<td>8: for ( i = 0; i &lt; R.Count; i++ ) do</td>
<td></td>
</tr>
<tr>
<td>9: if ( R[i] = L ) then</td>
<td></td>
</tr>
<tr>
<td>10: if ( \text{distance &lt; travelledDistance &lt; distance + distance} ) then</td>
<td></td>
</tr>
<tr>
<td>11: ( \text{return } q_{\text{section}}(\text{Vehicle}_i, t) = 1; )</td>
<td></td>
</tr>
<tr>
<td>12: else</td>
<td></td>
</tr>
<tr>
<td>13: ( \text{return } q_{\text{section}}(\text{Vehicle}_i, t) = 0; )</td>
<td></td>
</tr>
<tr>
<td>15: end if</td>
<td></td>
</tr>
<tr>
<td>16: ( \text{distance} += R[i].\text{distance}; )</td>
<td></td>
</tr>
<tr>
<td>17: end for</td>
<td></td>
</tr>
</tbody>
</table>

Algorithm 1 uses the enter time and exit time of each vehicle record, and estimates the average speed (Line 5) for the whole trip. By given the time \( t \), this algorithm can figure out in which road section the vehicle is driving, and it only returns one if this section is the same section we are analyzing (Line 11). The \( \text{FindPath}(O, D, \text{Vehicle}_i) \) (Line 1) is an function for estimating the path between O and D and it returns a set \( R \) of sections and their distances in meters. This method takes the following factors for consideration:

1. The shortest path: For a highway network, in contrast to other compact networks (such as urban road network), the shortest path is the primary choice for most drivers since there are not many options for alternative paths and an alternative path usually has a high time cost.

2. The OBU data: In addition to the toll data, we have OBU (On Board Unit) data which is collected from the OBU detector installed at certain road. The detectors are capable of capturing the MAC addresses and time information of the OBU device equipped on vehicles. By analyzing the OBU data, we can reroute some vehicles to the correct paths.

2) Algorithm 2: Calculate Traffic Density based on An Improved Section Travel Time Estimation Mode

The above Algorithm 1 is using the average speed mode to determine whether the vehicle is located on a dedicated section at any given time. Practically, the individual vehicle's speed varies, particularly restricted by the closed traffic. The vehicles usually tend to form a traffic “cluster” with the same or similar speed. To match this observed pattern, we propose an algorithm for our study that performs the iterated calculation of the traffic density, which redistributes the individual vehicles' travel speeds on each section by following the objectives below:

1. The total summation of variance for individual travel time on each section is minimized to a predefined threshold \( \theta \) after several iterations as below Equation 3, where \( T_{\text{section}} \) is the total vehicle travel time on the section. Therefore the speed of vehicles on each section at same time is as stable as possible.

   \[ \min(\sum_{\text{section}} \text{Var}(T_{\text{section}})) \leq \theta \]  

(3)

2. The maximum speed of individual vehicles is less the 120 kilometer/hour, which is the speed limit in the studied highway network.

   In each iteration, the mean value for each section travel time \( T_{\text{section}} \) is recalculated at first. Then, we use the individual section travel time derived from previous iteration \( T_{\text{section}} \), and redistribute the section travel time of each individual vehicle so that the variances (derived from mean time \( T_{\text{section}} \) and all \( T_{\text{section}} \) ) of each section can be balanced. Thus, instead of the average speed mode, this algorithm reasonably estimate the vehicle travel speed on each section based on the mean speed of other vehicles traveling on the same section. The pseudo code is described in Algorithm 2.
When \( Q_{ij} \) is identified, we can define the instant density for each section in Equation 4.

\[
\text{Density}_{i}(t) = \frac{Q_{i}(t)}{l 	imes n}
\]

By combining all the density values in a time sequence using Algorithm 2, we can depict the daily traffic density graph. Figure 1 shows a daily traffic density sample of an inbound road section (XiSanQi to BeiAnHe section) of Qing-zhang highway in Beijing network on 3rd June 2013. We take the measurement of section density every 15 minutes, and thus, there are 96 results. We observe that the density is relatively low during 12:00 - 6:00 AM and it reaches the maximum value (83) at 8:30 AM. Besides, the second peak value is 49, which is achieved at 5:15 PM. This observation indicates this is a typical weekday travel mode, and the traffic density during night time is at the lowest. The density achieves its peaks at both morning and afternoon.

### Congestion Index at the Section Level

To define the congestion index for each highway section, we normalize the traffic density into a range of [0,10) using the following logistic function in Equation (5):

\[
S(x) = \frac{1}{1 + e^{-0.05x - 0.5}} 	imes 20
\]

The logistic function [8] (illustrated in Figure 2) has been widely adopted as normalization function and it smoothly maps real values into the values in a bounded interval.

Then, the congestion index on each section can be expressed as in Equation (6):

\[
CI \_section(t) = S(\text{Density} \_section(t))
\]

According to the China Highway Service Standard [9], the LOS can be classified into 4 levels divided by the density. In our case, we quantify the congestion status for each road section into a value range from 0 to 10. Table II, next page, shows a conversion between the LOS defined in China highway and congestion index definition in our study.
TABLE II A CONVERSION BETWEEN THE LOS DEFINED IN CHINA HIGHWAY SERVICE STANDARD AND CONGESTION INDEX DEFINITION IN OUR STUDY

<table>
<thead>
<tr>
<th>Density PC/Km/Ln</th>
<th>China highway LOS</th>
<th>Congestion index</th>
<th>Congestion condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 7</td>
<td>Level 1</td>
<td>0 – 1.73</td>
<td>Free flow: Traffic flows at or above the posted speed limit and motorists have complete mobility between lanes</td>
</tr>
<tr>
<td>7 – 18</td>
<td>Level 2</td>
<td>1.73 – 4.22</td>
<td>Stable flow: The ability to maneuver through lanes is noticeably restricted and lane changes require more driver awareness. When reach the upper bound, Speed reduced and drivers are less comfortable.</td>
</tr>
<tr>
<td>18 – 25</td>
<td>Level 3</td>
<td>4.22 – 5.55</td>
<td>Unstable flows: Approaching the road capacity, the traffic flow becomes irregular and the speed varies drastically, because there is no way to maneuver in the traffic stream and 80% of the traffic is delayed.</td>
</tr>
<tr>
<td>25 – 45</td>
<td>Level 4</td>
<td>5.55 – 8.09</td>
<td>Forced or broken down flow. Every vehicle moves in the lock step. Travel time cannot be predicted, and the traffic volume exceeds the road capacity. A road in a constant traffic jam is at this LOS.</td>
</tr>
<tr>
<td>≥ 45</td>
<td>Level 5</td>
<td>8.09 – 10</td>
<td>Highly congested with long waiting queue</td>
</tr>
</tbody>
</table>

By applying the logistic function in Equation (5), the congestion index diagram for the same road section (XiSanQi to BeiAnHe section of Qing-Zhang highway) on 3rd July 2013 is illustrated in Figure 3. According to Table 2, during the morning hours for the day, the congestion index is over 8 until 9:30 AM, which implies that there is a long queue of waiting vehicles over this particular road section.

C. Congestion Index at the Road Level

A road typically includes a group of adjacent sections along a single direction. In the highway system, a road includes the arterial road which delivers traffic from collector roads to freeways, the ring road or beltway that encircles an urban area. Normally a road is crossing distinct regions, and thus there is a demand for understanding the road-level traffic status from the traffic administrator perspective. In order to identify the congestion index in a road which are joined by sections, we assign a weight value for each section against the entire road network. The section's weight value is measured with consideration of both topological importance and traffic volume. Using the weight ratio, we can then develop the road level congestion by merging these section-level congestion indices together.

1) Section Weight Evaluation

The section importance is introduced for the following reasons:

(1) The more traffic is affected by a road disruption; the greater economic lost is caused in transportation system;

(2) A damage on a frequently-used path will lead to a great travel cost since it is critical to most drivers' driving habits, and there are very few alternative paths offered in a highway network.

Therefore, the importance of an individual section \( e \) is measured by a hybrid combination by considering the level of flow volumes and the topological position of paths generated from the real traffic over a period of time:

\[
\text{Importance}(e) = \frac{\sum_{e \in R_{O,D}} (E_{R_{O,D}})}{\sum_{R_{O,D}} (E_{R_{O,D}})}
\]

Where \( O,D \) represents the origin and destination of each individual vehicle running within the observation period; \( E_{R_{O,D}} \) is the number of sections that the shortest O-D path \( R_{O,D} \) contains.

2) 3.3.2. The Road-Level Congestion Index

Since we can define the importance weight of each individual section \( e \) in the road network, the road-level congestion index for a particular road \( g \) at a given instant time \( t \) can be expressed as below in Equation 8:

\[
CI_g(t) = \sum_{e \in R} (\text{Importance}(e) \times CI_e(t))
\]

Figure 4 Error! Reference source not found., and Figure 5 demonstrate samples of 24 hours road-level congestion on major roads of Beijing highway during a weekday (3rd June 2013) and a weekend (8th June 2013). The section-level congestion index is calculated in every 15 minutes based on highway toll data. Then, by using the section importance (developed using toll data from January - June 2013), we obtain the congestion index for each road. These plots represent different roads in outbreak (green line) and inbound (blue line) directions for these arterial roads, as well as clockwise (green line) and anti-clockwise (blue-line) for the ring-road, respectively. The x-axis shows the time in hour and the y-axis shows the congestion index. Starting from the top-left corner clockwise, they are Jing-Zang arterial road, Jing-Cheng arterial road, Jing-Ping arterial road, Jing-Ha arterial road, Jing-Jin arterial road, Jing-Kai arterial road, Jing-Shi arterial road and the 6th ring road.
SHI FANG et al: TRAFFIC CONGESTION ANALYSIS USING HIGHWAY O-D TOLLGATE DATA

Figure 4. A sample of 24-hour congestion index for Beijing major roads on 3rd June 2013 (weekday). For each plot, the x-axis shows the time in hour and the y-axis shows the congestion index.

Figure 5. A sample of 24-hour congestion index for Beijing major roads on 8th June 2013 (weekend). The x-axis shows the time in hour and the y-axis shows the congestion index.

From both figures, we can observe that: on 3rd June 2013, except Jing-Cheng and Jing-Ping arterial roads, other roads are highly congested (with a congestion level above 8); and most of the highest congestion values are incurred in the afternoon. The overall congestion does not decrease in the weekend (8th June 2013). The congestion of inbound and outbound directions shows a similar variation pattern except the Jing-Zang (top-left corner) and Jing-Jin (bottom-right corner).

IV. A CASE STUDY ON CONGESTION PATTERN USING K-MEANS CLUSTERING

The identification of the congestion patterns helps researchers understand the nature of traffic flows. In many studies results indicates that there is a set of signature traffic patterns can collectively represent the ongoing traffic flow on each road of entire network [10]. In this case study, we examine the toll data of 134 days from 15th May 2013 to 25th September 2013, remove the data of public holidays within the period, and identify a number of different congestion patterns by using the K-means clustering method.

Finding a best fit K value for K-means is always a challenge. We use the silhouette method [11] to validate the clustering results for different K values. The silhouette method provides a succinct graph that can determine how well the clustering is performed. The silhouette value $S_i$ is calculated as

$$S_i = \frac{(b_i - a_i)}{\max(a_i - b_i)}$$

where $a_i$ is the average distance of $i^{th}$ point to the rest points in the same cluster, and $b_i$ is the lowest average distance of $i^{th}$ point to any other cluster. The clustering performance is determined by the closeness between $S_i$ and one. A sample of silhouette graph result is illustrated in Figure 6. This figure indicates that:

when there are two clusters, the silhouette value $S_i$ of the second cluster is close to one. The silhouette value that is close to one means the cluster is not overlapped with others, and the elements in this cluster are similar. The silhouette diagram also indicates there are a small number of negative points (at middle) and they are not considered as a cluster.

By examining the silhouette values from different clustering results, we can identify the best K value accordingly.

Figure 6. A sample of silhouette graph.

All arterial roads and ring-roads of Beijing network are studied individually in the clustering analysis. First, the daily road-level congestion values are inserted into $96 \times 1$ vector following the time line (the time granularity is on a basis of 15 minutes). Then the K-means clustering method is used with different K values and evaluated with the silhouette validation. The distance is calculated based on the Pearson
correlation coefficient since this study is focusing on finding different patterns of the congestion variation.

From the study, we found the clustering shows the best performance result (i.e., the silhouette value is the greatest) when only two or three road congestion patterns are separated from all roads in Beijing network. We selected two major arterial roads (Jing-Ping and Jing-Cheng) in this paper as demonstration.

In Figure 7, the clustering gives the best performance results when the Jing-Ping arterial road congestion patterns are separated in two clusters. The pie chart represents how the percentage of each congestion pattern is distributed within the seven days of a week. For example, in the right pie chart, the green color means 38% of the first congestion pattern appears on a Thursday. This pie chart implies that most of this pattern (88%) occurs on a weekday. There are 102 out of total 134 days when the pattern is found. The clustering result of Jing-Ping road congestion pattern clearly suggests that, in a weekday (right plot), this road has its congestion value peaked in morning and afternoon in both inbound and outbound direction. However, in a weekend day (left plot), the outbound direction (green color) congestion value achieves its peak at 10 am, and the inbound direction (blue color) congestion value has a peak at 4:30 pm. Meanwhile, both directions have only a single peak value.

Figure 7. The K-means clustering congestion pattern result for Jing-Ping arterial road.

In Figure 8, the Jing-Zang arterial road congestion patterns are divided into three clusters using K-means. The congestion patterns on weekend (center plot) shows single afternoon peak in inbound direction and single morning peak in outbound. However, the weekdays are separated into two modes (left plot and right plot). In contrast to the right plot, the left plot has a greater value of congestion for inbound during the daytime and an additional peak at 10 pm.

Figure 8. The K-means clustering congestion pattern result for Jing-Zang arterial road.

V. CONCLUSION

Researchers have conducted extensive studies on the road congestion evaluation using different strategies. In this paper, we introduce a new strategy to evaluate the congestion index of each highway road section of the entire network based on the real-world data. This congestion index measures the traffic intensity on both the section-level and road-level, and it can be applied directly in any different toll highway network. A case study on the daily congestion pattern analysis for Beijing highway was carried out, wherein a number of interesting traffic patterns have been identified.

REFERENCES