Spatial-Temporal Traffic Flow Pattern Identification and Anomaly Detection with Dictionary-based Compression Theory in a Large-Scale Urban Network

Zhenhua Zhang\textsuperscript{a}, Qing He\textsuperscript{b1}, Hanghang Tong\textsuperscript{c}, Jizhan Gou\textsuperscript{d}, and Xiaoling Li\textsuperscript{e}

\textsuperscript{a}Department of Civil, Structural and Environmental Engineering
The State University of New York, Buffalo, NY 14260
Email: zhenhuaz@buffalo.edu

\textsuperscript{b}Department of Civil, Structural and Environmental Engineering and
Department of Industrial and Systems Engineering
The State University of New York, Buffalo, NY 14260
Email: qinghe@buffalo.edu

\textsuperscript{c}School of Computing, Informatics, and Decision Systems Engineering
Arizona State University
Tempe, AZ 85281
Email: Htong6@asu.edu

\textsuperscript{d}Serco Inc., Reston VA 20190
Email: Jizhan.Gou@serco-na.com

\textsuperscript{e}Northern Region Operations,
Virginia Department of Transportation, Fairfax, VA 22030
Email: Ling.li@vdot.virginia.gov

\textsuperscript{1} Corresponding Author
Abstract
Traffic flow pattern identification, as well as anomaly detection, is an important component for traffic operations and control. To reveal the characteristics of regional traffic flow patterns in large road networks, this paper employs dictionary-based compression theory to identify the features of both spatial and temporal patterns by analyzing the multi-dimensional traffic-related data. An anomaly index is derived to quantify the network traffic in both spatial and temporal perspectives. Both pattern identifications are conducted in three different geographic levels: detector, intersection, and sub-region. From different geographic levels, this study finds several important features of traffic flow patterns, including the geographic distribution of traffic flow patterns, pattern shifts at different times-of-day, pattern fluctuations over different days, etc. Both spatial and temporal traffic flow patterns defined in this study can jointly characterize pattern changes and provide a good performance measure of traffic operations and management. The proposed method is further implemented in a case study for the impact of a newly constructed subway line. The before-and-after study identifies the major changes of surrounding road traffic near the subway stations. It is found that new metro stations attract more commute traffic in weekdays as well as entertaining traffic during weekends.

Keywords
Traffic anomaly detection; Traffic flow pattern identification; Dictionary-based compression theory;

1 Introduction
Studies on traffic flow patterns within a scale of road facilities have aroused increasing attentions in recent years. The traffic flow patterns can be taken as those characteristics of vehicle groups passing a point or a short segment during a specified span or traveling over longer sections of highway (Lan et al., 2008). The spatiotemporal features of traffic flow, occupancy, and speed over different time scales can provide insights into traffic operation and control, urban planning, incident management, etc. The time-of-day and day-of-week features of traffic flow or occupancy reveal the fluctuations of the jammed conditions and traffic operations of road links, intersections or even networks and can serve different research purposes. For instance, White et al., (2007) focused on the impact of the daily visitor transportation on the public infrastructures and suggested approaches to improve the environmental sustainability of national parks; Ramaswami et al. (1996) studied the network traffic patterns to design the logical topology and the routing algorithm so as to minimize the network congestion; Cassidy et al. (1999) investigated the characteristics of freeway traffic flow patterns and their findings have practical implications for freeway traffic planning and management; Lee et al.(2014) proposed a method to identify the congestion patterns from the big traffic data and predicted when the road congestion events will dissipate. Of all these studies, despite their different emphasis and
purposes, the traffic pattern identification usually performs as an initial step in the entire research process in transportation management, and there is still research gaps to explore further.

First, the measures for traffic flow patterns are not limited to traffic flow measurement. The metrics to describe the patterns usually vary from traffic flow (Cassidy and Bertini, 1999; Shen and Zhang, 2009; Zhang et al., 2001), density (Cassidy and Mauch, 2001; Treiber and Kesting, 2012), speed (Banaei-Kashani et al., 2011), etc. This is mainly due the feature of each metric being unique and the researchers usually have different purposes that their problems intentionally shape the definition of the traffic flow patterns. Besides traffic-related metrics, other information such as time and link locations may further enrich and clarify the connotations of pattern identification. For example, studies show that given the combination of direction, connectivity and locality of a road segment (and without having access to the actual traffic flow in the segment), one can distinctively determine the corresponding traffic signature (pattern) of a road segment with high probability (Banaei-Kashani et al., 2011). The features of the traffic flow patterns should be further explored combining both traffic, time and location information.

Second, traditional studies focused mainly on traffic flow patterns within intersections (Teodorovic et al., 2001) or corridors (Lan et al., 2008; Schoenhof and Helbing, 2007). These may not be sufficient to explain the pattern changes in a large-scale road network with hundreds of intersections. In recent years, the size of traffic data booms both in time and space, and the large-scale floating and fixed sensors are extensively utilized to collect the traffic data. The studies of pattern identification can expand from links to a district or even to a county. This expansion brings more difficulties and has aroused increasing attentions. There have been efforts in reducing the dimensionality of the traffic flow pattern problems, such as functional data analysis (Guardiola et al., 2014), principle component analysis (Jin et al., 2008). How to archive and summarize massive historical data effectively and extract meaningful traffic flow patterns from accumulated data to support decision making has become a significant challenge, considering the huge size of the dataset (Xu et al., 2013).

Third, traditional studies of traffic patterns drew their conclusions from different spatial and temporal perspectives including the day-to-day features (Hu and Mahmassani, 1997) or time-of-day features (Anbaroglu et al., 2014). The temporal traffic patterns obtained from traffic time series studies can reveal the pattern features in a different perspective (Li et al., 2015). Of all these studies, there is a pressing need to develop a systematic approach for traffic flow pattern identification and anomaly detection that consider both different spatial and temporal perspectives. The temporal features of the traffic flow pattern, e.g. at an intersection, should be unveiled by comparing with its surrounding traffic as well as its historical ones. Instead of combining “time” as a traffic-related variable, we proceed with a temporal perspective to study the changes of traffic flow patterns over time.
To address the above research gaps, we propose a method based on the compression theory in regional traffic flow pattern identification. Compression-based approaches have been successfully implemented in pattern recognition and anomaly detection in different domains, such as image processing (Akoglu et al., 2012), system query processing (Binnig et al., 2009), etc. Rather than trying to compress the set of frequent items, compression-based approaches compress the database and search for the subset of all frequent item sets that compresses the database best (Siebes et al., 2006). The method can capture the best regularities of the data with as little redundancy as possible and thus can capture the most important patterns in the datasets (Tatti and Vreeken, 2008). Previous research proves that this approach provides accuracy on par with the state-of-the-art in anomaly detection (Smets and Vreeken, 2011). Some even argue that this approach is competitive or superior to many of the state-of-the-art approaches in anomaly and interestingness detection, classification, and clustering with empirical tests on time series, DNA, text, XML, and video datasets (Keogh et al., 2007). Similar to the data fusion which can maximize the utility of the available (traffic) information (Treiber and Kesting, 2013), the compression method is capable of recognizing the frequent traffic flow patterns through effective interpretations of multi-dimensional data. Based on the compression theory and corresponding algorithm, we will be able to quantify the traffic performance in a given location and detect the abnormal traffic flow patterns from our studied network.

Figure 1 Three geographic scales to identify both the spatial and temporal traffic flow pattern within a road network
Through our study, we unveil the important features of traffic flow patterns separately from spatial and temporal perspectives. From the spatial perspective, the problem is how to measure and evaluate the anomaly degree of a road link or intersection approach as compared to others. This relates to the identification of recurrent traffic flow patterns and provides insights in how to constantly identify problematic locations within a certain scale of the road network. From the temporal perspective, the problem is how to measure and evaluate the abnormal time-of-day traffic flow patterns as compared to historical records. This relates to the non-recurrent traffic flow patterns and helps traffic operators learn why traffic goes to the extreme in one day whereas those in other days are relatively normal. What’s more, both spatial and temporal patterns can be interpreted in different geographic and time scales. We include three different geographic scales: sub-region level, intersection level, and detector (or lane) level, as illustrated in Figure 1. The traffic flow patterns in different levels of geography are further studied to explore the hourly and daily features of the traffic flow patterns.

The contributions of this paper lie in: First, we propose to employ compression theory to effectively interpret the large collections of multi-dimensional traffic data. The study area and method are fully detailed in Section 2 and 3; Second, we reveal the geographic distribution features, time-of-day features of traffic flow patterns through spatial and temporal traffic flow pattern identifications that are in Section 4; Third, physical meanings of the anomalies and comparisons with other methods are validated in Section 5; Findings of our method are concluded in Section 6 with a series of thoughtful discussions.

2 Data description
Our study builds on massive datasets collected by traffic loop detectors in the urban network of Northern Virginia (NOVA), located to the northwest bank of Potomac River and adjacent to the District of Columbia. The land area is more than 2500 sq. km including 3 major counties in Northern Virginia involved in this study: Loudoun County, Fairfax County and Prince William County. The area is further manually divided into more than 21 sub-regions by the Virginia Department of Transportation (VDOT) for traffic operation purposes, shown in Figure 2. Furthermore, the area holds nearly 2 million which is about one-quarter of the entire Virginia State and has long been known for its heavy traffic (Cervero, 1994).

We only focus on urban streets, including arterials, collectors and local roads, whereas freeways are not included in the scope of this study. More than 15000 traffic detectors are located on the links of over 1200 signalized intersections. All the detectors are lane-based stop-bar detectors, as shown in Figure 2(b). Therefore, the information collected by the detectors not only describe the traffic conditions on certain links but also those of the signalized intersections. The study period contains 8 months from January 1, 2014 to August 31, 2014. The raw data collected include the
traffic volume and occupancy in 15 minutes. Both traffic flow and occupancy are widely accepted in studying features of road networks or their impacts on daily traffic, since they can indicate other traffic flow characteristics (e.g. headway, speed, etc.) (Guardiola et al., 2014; Kerner and Rehborn, 1996; Kim et al., 2004). It is worth mentioning that some of the loop detectors may experience different kinds of malfunctions, so data is not completely trustworthy. The entire dataset has been carefully examined, and malfunctioning detectors are excluded from this study.

3 Methods
This section briefly discusses the feature categorization, compression methods, and Minimum Description Length. In addition, we derive an anomaly degree to quantify the anomaly levels of traffic flow patterns in either an intersection or a road network.

3.1 Feature categorization
The first task is to discretize the traffic-related features and for each feature and put the traffic data describing the similar traffic conditions into the same category based on its physical meanings. The features we need include traffic flow, traffic occupancy or other geographic information and categorization process requires the pre-definition of the bounds of the features. For geographic information, we studied the county, sub-region and intersection level all of which are discretized information. The county that a detector belongs to has three categories: Loudoun County, Prince William County, Fairfax County. The sub-region information is what we mentioned in Section 2. As to the intersection level, there are two categories: “Major” and “Non-Major”. The “Major” intersections are those whose major roads are arterials, and the “Non-Major” intersections are those whose major roads are collectors or local roads.

For the traffic information, the traffic flow per lane is categorized into three levels: Level A: [0, 700), Level B: [700, 1200), Level C: [1200, 2000] according to the service flow rate (Aashto, 2001; FHWA, 2000). The unit is vehicles per hour (vph). As the peak-hour link flow on a typical arterial may not exceed a certain limit, we exclude these data with flow higher than 2000 vph. It is worth mentioning that some studies classified the traffic flow into four stages: blocking flow, crowded flow, steady flow and unhindered flow (Rong et al., 2013). This is almost the same when the first two groups are merged into one. Given that the occupancy is linearly related to the density, the categorization should not only refer to (Aashto, 2001; FHWA, 2000) but also the study of the flow-density relationship in the past few decades. All of the previous studies accordantly indicate that same volume of traffic flow may correspond to twofold different density or occupancy values. Then, according to the definition of the level of service (LOS) from A to E, the traffic occupancy is categorized into five different levels: Level 1: [0, 1/17), Level 2: [1/17, 1/9), Level 3: [1/9, 1/7). After occupancy reaches 1/7, the traffic flow capacity is almost reached, and we introduce another 2 categories to distinguish 2 degrees of traffic jam: Level 4: [1/7, 1/2), Level 5: [1/2, 1]. After conversion, both the traffic flow and occupancy that are
collected in the 15-minute interval have been discretized. The discretization is the initial and also the crucial part of regional traffic flow pattern identification. The boundaries of each level in a category should follow the directions of level of service (LOS) (Aashto, 2001; FHWA, 2000) because it is widely accepted that traffic data falling into the same LOS follow the same traffic flow pattern. Level setting is crucial: if we set fewer levels to a category, the traffic flow patterns can not be properly distinguished. If we set more levels to the category, some of the free-flow traffic data may be identified as anomalies by the compression method because sometimes near-zero flows could be rare in a congested road network.

The geographic and traffic information constitute a database as the input of our algorithm. Table 1 briefly lists all the traffic-related features in our paper.

Table 1 Feature table

<table>
<thead>
<tr>
<th>Features</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>County</td>
<td>Loudoun County; Prince William County; Fairfax County;</td>
</tr>
<tr>
<td>Sub-region</td>
<td>Divided into 22 as shown in Figure 2</td>
</tr>
<tr>
<td>Intersection level</td>
<td>Major; Non-major;</td>
</tr>
<tr>
<td>Traffic flow</td>
<td>[0, 700); [700, 1200); [1200, 2000);</td>
</tr>
<tr>
<td>Traffic occupancy</td>
<td>[0, 1/17); [1/17, 1/9); [1/9, 1/7); [1/7, 1/2); [1/2, 1];</td>
</tr>
</tbody>
</table>

For the time information, we do not include any of them as features of the database. Instead, we try to aggregate the database according to the different time scales to identify the pattern differences in different time periods. Furthermore, we also consider the patterns on weekdays and weekdays separately in section 4.

3.2 Dictionary-based compression

After categorization, a raw database ending up with a rather “redundant” set of patterns has been created. The word “redundant” refers to that the database can be further compressed and generate useful knowledge. To exemplify the dictionary-based compression technique, we continue with a simple illustration. We assume to have a database \( D \) with 3 different features: Flow (\( F \)), Occupancy (\( O \)), Intersection (\( I \)). In each feature, there is a domain of possible values: \( \text{dom}(F) = \{F_1, \ldots, F_3\} \), \( \text{dom}(O) = \{O_1, \ldots, O_5\} \), \( \text{dom}(I) = \{I_1, \ldots, I_3\} \). The combination of all features in each 15-minute time period is taken as a database (DB) pattern: \( DF_i = \{F_j, O_k, I_l\} \). Theoretically, the domain of \( DF_i \): \( \text{dom}(DF_i) \) is all the possible combinations of features of the database.
The next step is to build a suitable pattern table \((PT)\) to compress and encode the features of the database. Table 2 shows an illustrative example of the Database Table and its PT. There are two columns in a PT. The first column is the PT pattern column. The domain of PT pattern: \(\text{dom}(PF)\) can be different from that of DB pattern \(\text{dom}(DF)\). The PT patterns can be the combination of any feature values but all PT patterns are included in \(\text{dom}(DF)\): \(\forall PF \in \text{dom}(DF)\). The second column is the number of occurrences of each PT pattern in the database. The PT then performs as a code dictionary and the process of converting the DB patterns into combination of PT patterns is called dictionary-based compression. One can see that after conversion, the previous larger database are compressed and encoded into a smaller one and the usage of PT patterns differs from each other in the PT Table. One can assume that different pattern tables result in different PT patterns and thus different usages.

<table>
<thead>
<tr>
<th>Database table</th>
<th>Pattern table</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB pattern ((DF_i))</td>
<td>PT pattern included</td>
</tr>
<tr>
<td>F</td>
<td>O</td>
</tr>
<tr>
<td>Level A</td>
<td>Level 2</td>
</tr>
<tr>
<td>Level A</td>
<td>Level 2</td>
</tr>
<tr>
<td>Level A</td>
<td>Level 2</td>
</tr>
<tr>
<td>Level A</td>
<td>Level 2</td>
</tr>
<tr>
<td>Level A</td>
<td>Level 2</td>
</tr>
<tr>
<td>Level A</td>
<td>Level 2</td>
</tr>
<tr>
<td>Level B</td>
<td>Level 4</td>
</tr>
<tr>
<td>Level B</td>
<td>Level 3</td>
</tr>
</tbody>
</table>

For DB pattern encoding, given the usage, one can compute the optimal lengths of the code words to encode the patterns according to optimal prefix code (Rissanen, 1978). The length of PT pattern \((PF_i)\) in a certain pattern table is defined as:

\[
L(\text{usage}(PF)|PT) = - \log \left( \frac{\text{usage}(PF)}{\sum_{PF_i \in PT} \text{usage}(PF_i)} \right)
\]

\(1\)

It is worth mentioning that the base of all logarithms in this paper is 2 (Agrawal and Srikant, 1994). Then, the length of DB pattern \((DF_i)\) is calculated as the sum of the lengths of all PT patterns it contains. We find the best set of \(PF\) that can make up \(DF\). In one \(DF\), any component \(PF\) cannot cover other \(PFs\).
\[ L(DF|PT) = \sum_{PF \in DF} L(usage (PF)|PT) \]  
\[ L(DB|PT) = \sum_{DF \in \text{dom}(DF)} L(DF|PT) \]

The length of DB is the sum of the lengths of all DB patterns the database contains.

For encoding the pattern table, we still need optimal prefix code (Akoglu et al., 2012). There are two parts involved in the length of PT. The first part is the sum of lengths of all PT patterns; the second part is the sum of lengths of all singleton items in each category in DB. Define \( l \) as all the singleton items in DB, \( c \) as the total count of singleton items and \( r_i \) is the count of the \( i \)th singleton item. For example, in Table 2, \( c \) is equal to 24, \( r_i \) of the singleton item “Level A” is 6.

The length of the PT table is defined as:

\[ L(PT) = \sum_{PF \in PT} L(usage (PF)|PT) + \sum_{r_i \in I} -r_i \log \left( \frac{r_i}{c} \right) \]

The length of a DB pattern code plays a significant role in anomaly detection. To save storage space, patterns that occur more frequently result in smaller \( L(PT) \), which are regarded as the normal ones, whereas those occurring less frequently are taken as the abnormal ones. For example, in transportation study, the unexpected non-recurrent traffic congestions are treated as anomalies and the severity of the anomalies can be quantified by the length of the DB patterns.

### 3.3 Minimum Description Length principle

One can see from Section 3.2 that given a database of observations, the lengths of both DB and PT are totally decided by the selection of PT. One principle, called Minimum Description Length (MDL), should be followed to select a suitable PT to compress and encode the database. The MDL principle identifies the best PT which minimizes the description length:

\[ \mathcal{L} = L(DB|PT) + L(PT) \]

The MDL principle requires us to find the description length of the shortest codes for the actually observed sequence (data), rather than a mean length (Barron et al., 1998). Also, one can see that a complex PT with diverse PT patterns can compress the DB very well and thus leads to a small \( L(DB|PT) \). The side effect is that it will also result in a large \( L(PT) \). MDL principle attempts to balance the complexity of PT and its fit to DB. Therefore, the PT that can provide the shortest description length of Eq. 5 is the best pattern table that compresses the database.

To find the best PT Table, we employ a heuristic search algorithm as follows.

**Algorithm: Dictionary-based compression**
Input: Database with \( n \) rows and \( m \) categories

Output: A PT table and the length of each pattern

Build the initial PT table and all PT patterns \( PF_i \) are singleton items of features in DB

Compute the initial description length \( L_0 \), the optimal length \( L = L_0 \)

Implement the Apriori algorithm to find all frequent items \( FI \) whose frequency is higher than a threshold \( T \), these frequent items constitute a set \( S \)

Repeat

for \( FI_i \) in \( S \)

Put \( FI_i \) into the PT table

Compute the current description length \( L_i \)

If \( L_i < L \)

\[ L = L_i \]

remove \( FI_i \) from \( S \)

add \( FI_i \) into PT table

else

remove \( FI_i \) from \( S \)

until \( |S| = 0 \)

The proposed search algorithm can be interpreted in the following steps:

- Step 1: All possible singleton items in DB are taken as the PT patterns in the PT table. For example in Table 2, the PT patterns are (“Level A”, “Level B”, “Level 2”, “Level 3”, “Level 4”, “Major”, “Non-major”). According to these PT patterns, one can calculate the initial description length and take it as the current length.

- Step 2: We implement the Apriori algorithm (Agrawal and Srikant, 1994) to find all frequent items that are the combinations of one or more singleton items. We order the items according to their frequency and choose those whose frequency is higher than a threshold. These “frequent” items are the potential candidates for PT patterns. For example, in Table 2, the combination of “Level A” and “Level 2” has the highest frequency of 6 and they are put on the list of candidates.

- Step 3: We add the most frequent combination from the list of candidates into the PT table and recalculate the description length. If the recalculated length is smaller than the current one, we add the new item to the PT table and change the length of each pattern. If the recalculated length is larger, we keep the previous PT table and score. For example, in Table 1, after calculation, it is found that adding the item “Level A/Level 2” to the PT table can reduce the description length. Therefore, the item “Level A/Level 2” is chosen and added to the PT table.
• Step 4: We remove the item in Step 3 from the candidate list and continue with the next candidate until there is no candidate left on the list.

It is worth mentioning that the threshold $T$ for the candidate item is chosen as 30% of the total count of DB patterns in this paper. One can still lower the threshold but this may lead to more computation and relatively less improvement.

3.4 Anomaly degree
In Section 3.3, the MDL principle finds the best pattern table, and we proceed to derive a normalized anomaly degree that can characterize the traffic patterns.

Given a PT, the length of a traffic pattern $PF$ is:

$$L(DF|PT) = \sum_{PF \in DF} L(usage(PF)|PT)$$

$L(DF|PT)$ indicates the anomaly degree of a traffic pattern. The higher the length is, the closer the traffic pattern length is to the upper bound, and consequently the more abnormal the traffic pattern should be. Theoretically, when the DB table has only one kind of $DF$, there should also exist only one kind of $PF$ equal to $DF$ in $PT$. The lower bound of $L(DF|PT)$ is inevitably 0. However, the upper bound should be bounded by a certain value. We continue to approximate the upper bound by estimating the possible largest values of the pattern length.

**Definition 1:** Given the number of DB patterns, the approximate upper bound of $L(DF|PT)$ is defined as:

$$U = \tau \cdot \sum_{i \in dom(i)} - log \left( \frac{1}{N} \right)$$

Where $N$ is the total number of rows in DB, $\tau$ is a discount factor.

**Remark 1:**

$$L(DF|PT) = \sum_{PF \in DF} L(usage(PF)|PT)$$

$$= \sum_{PF \in DF} \max\{L(usage(PF)|PT)\}$$

$$= \sum_{PF \in DF} \max\left\{-log \left( \frac{usage(PF)}{\sum_{PF \in PT} usage(PF_i)} \right) \right\}$$

Two cases need to be considered separately.
Case 1: there exists only singleton items in PT table, then the highest values of 
\(-\log \left( \frac{\text{usage (PF)}}{\sum_{PF \in PT \text{ usage (PF)}}} \right)\) for each PT feature in DF should be usage (PF) = 1 and 
\(\sum_{PF \in PT \text{ usage (PF)}} = N\) that is 
\(-\log \left( \frac{1}{N} \right)\).

Case 2: there exist PT patterns that are combinations of 2 or more singleton items. For PT patterns with more than one singleton items, 
\(-\log \left( \frac{\text{usage (PF)}}{\sum_{PF \in PT \text{ usage (PF)}}} \right) = \n\sum_{i \in dom(i)} - \log \left( \frac{1}{N - \sum_{i \in dom(i)} M_i} \right)\) where \(M_i\) equals to the number of combined PT patterns that contain the singleton items in \(i\)th feature. Also, \(\sum_{i \in dom(i)} - \log \left( \frac{1}{N - \sum_{i \in dom(i)} M_i} \right) \leq \n\sum_{i \in dom(i)} - \log \left( \frac{1}{N} \right)\).

However, after our examination, the circumstance that all features with usage (PF) = 1 is really rare and \(\sum_{i \in dom(i)} - \log \left( \frac{1}{N} \right)\) is too large for most of DB table. To make the anomaly degrees from different DB comparable and also make it less conservative, we introduce \(\tau\) as a discount factor. Therefore:
\[ L(DF|PT) \leq \tau \cdot \sum_{i \in dom(i)} - \log \left( \frac{1}{N} \right) \]

In this paper, \(\tau\) is set as 0.85. Defining \(U\) allows us to normalize the anomaly index to be a fraction number. We further derive the anomaly degree to evaluate the performance of traffic patterns:
\[ D(DF_i|PT) = \frac{L(DF_i|PT)}{U} \]  \(8\)
Where \(D(DF_i|PT)\) is the normalized anomaly degree of the \(i\)th traffic pattern in the database. Note that since \(U\) is an approximated value, \(D(DF_i|PT)\) could possibly exceed 1.0.

One can see that the anomaly index is a comparable index and can identify the least frequent traffic flow patterns as the abnormal ones. This index is valid because given the vast study area in NOVA, the places that experience abnormal traffic operations such as traffic jam, traffic accident, festival parades etc. take up only a small portion of the whole area. This rule also applies to the abnormal time periods of a given location. Furthermore, from Eq. 5, this index includes more than one traffic-related metrics including the geographic information. Through compression, this inclusion of geographic information can decrease the index where the traffic is always jammed (abnormal) as well as increase those where the traffic is unexpectedly abnormal. It can be concluded that the anomaly index can compare the overall performance of traffic operations in different places or different time periods. The lower the value is, the less abnormal the traffic condition and the better traffic performance should be. The absolute value also poses significance on the traffic performance evaluation. Our empirical examination shows that most of the values are below 0.7. The reasons anomaly index higher than 0.7 may be another
interesting topic and not be covered in our study. What interests us is the performance measures in different spatial and temporal perspectives.

Here we define two anomaly index: spatial anomaly degree and temporal anomaly degree:

- For spatial anomaly degree, the DB table is the set of traffic flow patterns in the whole region over the same time period.
- For temporal anomaly degree, the DB table is the set of traffic flow patterns in one location over the same time period on different days.

One can see that both spatial and temporal anomaly degrees aim at capturing the traffic flow patterns that are different from the rest majority. The spatial anomaly degree may indicate locations that have much heavier traffic than their surrounding ones such as overcrowded traffic in afternoon peak, typically caused by recurrent commuter traffic. The temporal anomaly degree may indicate the time periods in one day when traffic are usually lighter or heavier than other days; This index is important to identify non-recurrent traffic flow pattern that may be caused by unexpected incidents vehicle breakdowns, crashes, the closure of road links, festival parades, malfunction of the traffic signals, etc. Previous studies show that the non-recurrent traffic flow pattern is even more important because it accounts for more than 50% of the total motorist delay (Chin et al., 2002).

The causes of traffic accidents are various and the corresponding strategies should be devised accordingly which may include traffic signal control (Asamoah, 2014; Ding et al., 2015; He et al., 2011, 2014), urban design (Saelens et al., 2003), work zone control (Beacher et al., 2004), incident management (He et al., 2013; Ozbay and Kachroo, 1999), etc. These will not be covered in our study. Instead, we focus on the anomaly detection method and its applications.

4 Numerical examples
In this section, we implement the proposed traffic flow pattern identification algorithm in NOVA area. We will first explore the physical meaning of the anomaly degrees by comparing results between our compression-based method and other pattern recognition methods, then reveal the spatial and temporal traffic flow pattern characteristics and finally prove the validity of our method by a case study.

4.2 Spatial regional traffic flow pattern
The spatial regional traffic flow pattern helps identify the abnormal locations that display different characteristics from the rest locations. It is a straight-forward tool for traffic operators to compare the traffic flow patterns in different locations over the same time period. We first calculate the anomaly degrees of the traffic flow patterns in the detectors and further derive the
anomaly degrees of the intersection by taking the 90th percentile of anomaly degrees of all
detectors that belong to the intersection. This is reasonable because the overall performance of
the intersections can be worsened by one or two jammed directional links. Figure 3 shows the
heat map of spatial regional traffic flow patterns at intersection level in PM peak (5:00-6:00
p.m.) separately on a weekday (Feb 20, 2014) and a weekend (Mar 22, 2014). The scale of the
color in different plots is set to be equal in order to make the anomaly degrees comparable. It is
concluded that the red (dark) dots indicate a totally different traffic flow pattern from the rest
majority and the redder the dots are, the more abnormal the places should be. The detector
anomaly degrees of two most abnormal intersections are displayed below the heat map.

Here we have several interesting observations. First, during PM peak, the overall anomalies
mainly concentrate in a few intersections, indicating the traffic jam and also the lack of road
facilities. One can see some distinct red corridors, such as Leesburg Pikes, which goes from
northwest to southeast. This indicates that certain roads in the network experience more severe
conditions than that of others. In addition, one can say that certain regions are more likely to be
jammed together (Banaei-Kashani et al., 2011). The numerical results of some studies revealed
that different traffic flow patterns could emerge depending on the (different) observed time-scale
(Lan et al., 2008). In our study, however, there is almost no difference in geographically
distributed features of anomaly locations on weekdays or weekends during PM peak.

To identify the day-to-day fluctuations of spatial traffic flow patterns, we randomly selected two
intersections in our studied region, which are located in different counties. The time period is
also set in PM peak (5:00-6:00 p.m.). From Figure 4, most of the traffic flow patterns do not
show large fluctuations from January to September. This means that the traffic performance of
spatial traffic flow patterns for one intersection over certain time periods in a network remain
almost constant in different days.
The unchangeable nature of the traffic flow patterns may not be true for those traffic flow patterns in one day during different time periods. Due to the fact that we have more than 1200 signalized intersections, aggregated anomaly degrees in different sub-region may be more helpful in identifying the pattern shift. Table 3 presents the anomaly degree results of three sub-
regions. The anomaly degrees increase as the geographic locations gradually move closer to the District of Columbia. Also, from Figure 5, one can see a clear time-of-day pattern shift in almost all sub-regions, a trend of regional anomaly degrees decreasing from AM peak to noon and increasing back to PM peak. In addition, the ranks of the sub-regions remain almost the same.

The spatial traffic flow pattern identifications and the corresponding anomaly detection may provide useful information for traffic operators. The places with higher anomaly degrees may experience heavy traffic and elongating travel time. As some studies on the congested traffic states that typical kind of traffic congestion depends on the specific freeway (Treiber et al., 2000), the abnormal traffic flow patterns also emerge on specific road lanes. The causations may be the layout of the road network or other infrastructure reasons and it should be an interesting topic in further studies.

Figure 4 Anomaly degrees of spatial traffic flow patterns of two intersections on PM peak (5:00-6:00 p.m.) in different days. The first intersection (a) is Richmond Hwy. and Sherwood Hall Ln.; the second intersection (b) is Lee Hwy. and ramp of W Ox Rd.
Table 3 Anomaly degrees of spatial traffic flow patterns in 3 sub-regions in different time periods on weekday February 2nd, 2014.

<table>
<thead>
<tr>
<th>Region ID</th>
<th>AM peak (7:00-8:00)</th>
<th>Noon (11:00-12:00)</th>
<th>PM peak (17:00-18:00)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.725</td>
<td>0.668</td>
<td>0.722</td>
</tr>
<tr>
<td>19</td>
<td>0.686</td>
<td>0.641</td>
<td>0.673</td>
</tr>
<tr>
<td>16</td>
<td>0.593</td>
<td>0.561</td>
<td>0.590</td>
</tr>
</tbody>
</table>

Figure 5 Spatial anomaly degrees in three different time periods for different sub-regions (each link represents a sub-region).

4.3 Temporal regional traffic flow pattern

The temporal regional traffic flow pattern aims to identify the abnormal patterns over certain time periods in one day that display different pattern characteristics from the other days. Same as the spatial traffic pattern identification, we first calculate the anomaly degrees of the detectors and aggregate them into the anomaly degree of the intersection by taking the 90th percentile. We also consider weekdays and weekends separately.

Figure 6 shows a quite different heat map of temporal traffic flow pattern than that of spatial traffic patterns. As one can see, there is no clear corridor pattern of abnormal intersections. The occurrences of abnormal traffic flow patterns are quite random. If the criterion of anomaly degree is set as 0.8, less than 10 percent of the intersections are abnormal, whereas more than 90% of the intersections operate pretty much the same as most of the other days. For two most abnormal intersections, the anomaly degrees of the detectors in the intersection differ greatly.
from each other, and the traffic flow pattern of an intersection may deteriorate by one or two abnormal detectors.

We further select two intersections and examine the temporal traffic flow pattern fluctuations over the PM period in different days. From Figure 7, we can see that the abnormal degree oscillates in different days. The rare peak points are identified as abnormal patterns, whereas most of the rest represent normal patterns. The temporal abnormal pattern could be observed during both weekends and weekdays and these can help the traffic operators identify the unexpected events in certain locations.
Figure 6 Temporal regional traffic flow pattern identification during PM peak (5:00-6:00 p.m.) on (a) Feb 20, 2014 (Weekday) and (b) Mar 22, 2014 (Weekend). The anomaly degrees of all detectors (x-axis of the bar plots) in two most abnormal intersections and their locations are shown in the bottom of the plot.
Table 4 shows the anomaly degrees of temporal traffic flow patterns in different sub-regions in different time periods on the same weekday. Unlike the spatial flow pattern, the temporal anomaly degrees in the sub-regions do not show a clear trend in different time periods as shown in Figure 8. This observation also proves that the temporal flow pattern has little to do with geographic locations.

One can see that the identification of temporal traffic flow pattern may find the unexpected anomalies that appear quite randomly in both time and space. It can be assumed that under this circumstance, the traffic operations or travel time over certain links may be greatly influenced.

Table 4 Anomaly degrees of temporal traffic flow patterns in 9 sub-regions in different time periods on the same weekday
<table>
<thead>
<tr>
<th>RegionID</th>
<th>AM peak (7:00-8:00)</th>
<th>Noon (11:00-12:00)</th>
<th>PM peak (17:00-18:00)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.220</td>
<td>0.178</td>
<td>0.187</td>
</tr>
<tr>
<td>19</td>
<td>0.225</td>
<td>0.175</td>
<td>0.227</td>
</tr>
<tr>
<td>16</td>
<td>0.331</td>
<td>0.205</td>
<td>0.212</td>
</tr>
</tbody>
</table>

Figure 8 Temporal anomaly degrees in three different time periods for different sub-regions (each link represents a sub-region).

### 4.4 A case study

Serving different purposes, the spatial and temporal traffic flow patterns in one location may display different features. These features can jointly explain the changes of traffic flow patterns in two different perspectives. Figure 9 illustrates anomaly degrees of all detectors over three time periods: AM peak (7:00-8:00 a.m.), Noon (12:00-13:00 p.m.) and PM peak (16:00-17:00 p.m.), across two days, Feb 20, 2014 (weekday) and Mar 22, 2014 (weekend). One can see that anomalies may coexist both in time and space even though the situation is very rare. The correlation between the spatial and temporal anomaly degrees is not high. This means the spatial and temporal anomaly may not possibly influence each other and should be explained separately. The spatial patterns compare the traffic flow patterns across the geo-spans, whereas the temporal ones compare those across time-spans. Thus, if a location is congested all the time as compared to other locations in the network, it is abnormal in spatial traffic flow pattern; but it is possible that these such congested conditions are only normal day-to-day recurrent pattern if compared with the historical traffic flow patterns during the same time-of-day. Therefore, the spatial anomaly does not necessarily indicate temporal anomaly and vice versa. Combinations of both traffic flow pattern identifications can enhance our understanding of the regional traffic patterns.
We implement the proposed method to study the changes of traffic flow patterns of the new 11-mile extension of Sliver Line, a subway line of Washington Metro. The extension consists of 5 exclusive new stations, which began service on July 26, 2014. It connects with the Orange Line at East Falls Church Station and extends to the west business district. 4 out of 5 stations are located in the densely populated area of Tysons Corner and may bring impact to the local traffic and the regional traffic flow pattern. One can refer to Table 1 for feature and pattern definitions used in this case study.

We calculate the anomaly degrees of spatial and temporal traffic flow patterns of all detectors near the metro stations separately and then take an average of the anomaly degrees for each intersection. To make a clear comparison, we further calculate the ratios of the degrees after and before the Silver Line extension. Still, we consider the case on weekdays and weekends.

Figure 9 Spatial and temporal regional traffic flow pattern of all detectors on Feb 20, 2014 (Weekday) (a) 07:00-08:00 a.m., (b) 12:00-13:00 p.m. and (c) 17:00-18:00 p.m. and Mar 22, 2014 (Weekend) (d) 07:00-08:00 a.m., (e) 12:00-13:00 p.m. and (f) 17:00-18:00 p.m.
separately. Figure 10(a) shows the geographic information of the sub-region containing these 4 metro stations. In the same geographic area, Figure 10(b)-(e) shows the ratios of spatial and temporal anomaly degrees before and after case separately on weekdays and weekends.
Figure 10 (a) Map of the sub-region; Ratios of spatial anomaly degrees between after and before Sliver Line extension in the same area (x-axis and y-axis are longitude and latitude, respectively) on (b) weekdays and (c) weekends; Ratio of temporal anomaly degrees between after and before Sliver Line extension on (d) weekdays and (e) weekends.

△ Metro stations  ● Signalized
It will be helpful to examine the anomaly degrees together with the geographic-distributed features of the study areas. From Figure 10(a), the metro stations are evenly distributed along the road line of Route 7 and Chain Bridge Rd and one of the metro station (Greensboro metro station) is almost located at the intersections of two roads. On both arterials, one can see an overall increase of anomaly degrees both in spatial and temporal traffic flow patterns on weekdays and weekends after the Silver Line extension. It shows that new metro stations attract more commute traffic and entertaining traffic (for Tysons Corner Center in the red area). Besides these two arterials, other intersections, especially those collectors and local roads located within the commercial land, do not show an identical increase or decrease in different spatial-temporal perspectives. Some places even show a clear decrease of anomaly degrees which means less traffic jam is observed. As a conclusion, the newly built metro station may change the spatial and temporal traffic flow patterns. Potentially it worsens the traffic on major arterials and alleviates that of collectors and local roads.

For the spatial traffic flow pattern part, it should be noted that the metro stations are coupled with several newly-built transit lines designed to connect new Silver Line rail travel (WMATA, 2014). Most of the transit lines stop at the arterials instead of collectors or local roads. What is more, new parking lots are open together with the metro station and their locations are mostly near the arterials. The newly-built metro station may change both the land-use features and the trip mode and in turn, change the spatial traffic flow patterns. Over a certain time period, some links become more crowded and others become relatively less. One can say new transit systems possibly change the recurrent traffic flow patterns within a geographic scale. For the temporal traffic flow pattern part, these changes are even more apparent. The after-case pattern can be taken as the minority and detected just because the time period of after-case pattern is from July 26th to August 31st which is only a small portion of the total record period. Unlike spatial patterns, temporal traffic flow patterns capture the changes of recurrent traffic patterns as with the changes of transportation systems (e.g. new subway stations). It is worth mentioning that with longer observation time periods, the recurrent pattern changes will become stable and the new patterns will be regarded as normal ones. In sum, the spatial and temporal traffic identification reach conclusions in different ways and support decision making for transportation planning and management.

5 Comparisons
5.1 Comparisons between different discretization settings
We further conduct a study with the proposed methods by different discretization boundary settings of flow and occupancy in calculating the temporal anomaly degree.

We present the results in a flow-occupancy plot, coming after the concept of the fundamental diagram (Geroliminis and Daganzo, 2008), which assumes that flow and density (occupancy)
follow a concave relationship. We randomly select one detector located on different links and conduct a temporal traffic flow pattern identification on these detectors. The DB is the set of traffic flow patterns (here only traffic flow and occupancy) over the same 15-minute period on different weekdays (weekend data are excluded).

Figure 11 Examples of anomaly degrees in one detector during 17:00~17:15 p.m. on different weekdays. Obvious abnormal dots are circled. The boundary settings are defined as follows, (a) the same setting as introduced in Section 3; (b) different levels when discretizing traffic flow to [0, 500), [500, 1000), [1000, 2000]; (c) different levels when discretizing traffic occupancy to [0, 20%), [20%, 40%), [40%, 60%), [60%, 80%), [80%, 100%].

As shown in Figure 11 (a)-(c), the darker red dots indicate more abnormal traffic flow patterns. One can see from Figure 11 (a) that the traffic flow is around 500~700 during the peak hours. As compared, those traffic flow data lower than 500 with small occupancy are more abnormal. This is reasonable because relatively low traffic flow and occupancy happen during a rush hour which makes the traffic flow patterns on those days quite different from other days. For those isolated traffic data whose occupancy is apparently higher than 0.15, the patterns are also taken as the anomalies. The observations by the compression method in Figure 11 (a) coincide with our intuitive perceptions based on the fundamental diagrams. The results based on the even discretization in traffic flow and occupancy are shown in Figure 11 (b) and 11 (c), respectively. In Figure 11 (b), with the upper bound for free flow is set to be 500, the anomaly degrees for those traffic flow patterns are relatively higher. In Figure 11 (c), the high upper bound (20%) for the first occupancy level already includes the majority of the traffic occupancy data. Manually changing the boundary settings may come up with totally different results and these results may not conform to the current LOS requirements.
Figure 12 (a)-(f) shows the effects of merging and dividing of original discretization levels on the anomaly degrees into two different detectors. The anomalies in two detectors can be explained from different flow and occupancy interpretation. In the first detector, the anomaly data are due to high occupancy while in the second one, those are mostly due to low traffic flow. Their results show that merging levels may decrease the anomaly degree values while dividing may do the reverse. This can be explained by the fact that a smaller number of levels corresponds to a smaller value of $L(usage \, (PF)|PT)$ in Equation 6 and in turn a higher value of anomaly degree. Merging and dividing the levels may over or underestimate the anomaly degrees but the ranks of each traffic flow pattern shall not change apparently. However, it can be assumed that if the size of the datasets is small, the effects may be quite unpredictable as smaller sample size may be less likely to uncover the regularities in the datasets.
Comparisons with another anomaly detection method

We further compare our proposed method with long-term anomaly detection method proposed by Vallis et al. (2014). The Vallis’ model is built on a daily basis and requires the detection of the trend component. It can detect anomalies according to the periodic features of the time-of-day traffic data. The method employs time series decomposition and robust statistics for...
detecting anomalies. It builds on the generalized Extreme Studentized Deviate test (ESD) (Rosner, 1983) which assumes data set comes from a normal-distributed population. According to Zhang et al. (2016), the traffic occupancy over a certain period may follow a normal distribution. Thus, the method can potentially detect the traffic flow pattern anomalies with high precisions. Figure 13 shows the results on the same dataset used in Section 4. For display purpose, we only present the results of the time period of 17:15 ~ 17:30 p.m. on different days. Both methods present different anomaly detection results making it worthwhile discussing. In contrast, compression-based method emphasizes on the features over certain time periods on different days.

Figure 13(b) and 13(c) present the data over the same peak period on different days. The anomalies detected by our method can be better explained. Two extremely low traffic flow and occupancy (shown as blue triangles) are taken as anomalies because they indicate traffic conditions that are substantially different from the majority of other peak periods. In contrast, Vallis’ model outputs three anomalies, which are shown as red rectangles. Two of three are with high flow and high occupancy, which are quite normal for peak periods. Therefore, temporal anomalies are not fully captured by Vallis’ model. Compared with Vallis’ model, the compression-based method finds these abnormal points partially or even wholly decided by the frequencies of their appearances. It opens a new window in detecting the abnormal traffic flow pattern through the massive collection of traffic-related data. One can see that the compression technique identifies the abnormal points from a different perspective. The effectiveness of these methods can be further explored for more specific purposes of traffic flow pattern identification.
Conclusions and discussion

As a first attempt, this paper leverages dictionary-based compression theory for regional traffic flow pattern identification and anomaly detection within a large-scale traffic network. Spatial and temporal anomaly degree indices are derived to describe and quantify the abnormal traffic flow pattern. Empirical results using traffic flow and occupancy show its adaptability in the

Figure 13 (a) Temporal anomaly results by the proposed compression-based method, (b) traffic flow and (c) traffic occupancy over time period 17:45 ~ 18:00 p.m. in different weekdays. Red rectangles indicate the anomalies detected by (Vallis et al., 2014); blue triangles indicate those detected by compression-based method.

6 Conclusions and discussion

As a first attempt, this paper leverages dictionary-based compression theory for regional traffic flow pattern identification and anomaly detection within a large-scale traffic network. Spatial and temporal anomaly degree indices are derived to describe and quantify the abnormal traffic flow pattern. Empirical results using traffic flow and occupancy show its adaptability in the
transportation field. There are three major advantages to this approach that should be highlighted:

- Effective reduction of dimensionality: two or more traffic-related metrics are interpreted together which prevents the complex definition of several different variables and parameters. The features of traffic flow patterns can be quantified as a low-dimensional index.
- Clear representations of geographic-distributed features of spatial and temporal traffic flow patterns: In the heat map of our study area, anomaly locations can be easily identified.
- Clear representations of time-of-day features of spatial and temporal traffic patterns: the time-of-day traffic flow patterns can clearly identify the pattern changes over the same time period on different days or different time periods in the same weekday or weekend.

The spatial pattern identification presents meaningful results of recurrent traffic flow patterns: Over a certain time period, the heat map of spatial pattern anomaly in the region shows almost identical results on weekday and weekend. Two most abnormal intersections in the plots show that if one detector suffers a higher anomaly degree, other detectors in the same intersection are more probable to be abnormal. Further studies on sample locations show that the spatial traffic patterns remain almost the same for different days. But if we focus on the traffic patterns in different time-of-day on the same day, an identical trend from AM peak, Noon, and PM peak can be found in different sub-regions.

The temporal traffic pattern identification demonstrates the non-recurrent features of traffic patterns. In our study area, the occurrence of temporal abnormal places are quite random, and detectors in the same intersections may have quite different anomaly degrees from each other. As compared to the spatial traffic pattern, temporal anomaly degrees in the same time period on different days usually have larger fluctuations and the higher value may indicate a bad traffic performance.

Also, it is proved that there exists almost no correlation between spatial and temporal anomaly degrees. The combination of the two leads us to a better understanding of traffic flow pattern. Results of a case study show the changes in the traffic flow patterns that are influenced by major municipal constructions over a certain scale of geography. The changes of land use and travel modes can, in turn, be identified as a shift in traffic flow patterns. By exploring the physical meanings behind the abnormal traffic flow patterns and comparing it with other state-of-the-art methods, one can see that the compression-based method identifies the abnormal points from a different perspective.
Although the proposed compression-based method is promising, its applications in transportation data analysis still show certain limitations. First, the method works better on a large dataset. It’s better to employ the method on the historical traffic data with long recording period. When dealing with small datasets, the results may not be convincing. Second, the method is very sensitive to outliers, and its application relies a lot on the credibility of the raw database. The researchers need to carefully prune the raw data before generating anomaly degrees. Besides these shortcomings, this method provides insight into the interpretation of the multi-dimensional traffic-related data, and some research questions remain, especially in incident detection, network evaluation, and traffic management. Also, as the geographic distribution of spatial traffic flow patterns remains almost the same in different time periods, there may exist room for improving planning or management of the road networks. Strategies for improving the traffic jams can be evaluated by comparing the before and after anomaly degrees. The temporal traffic flow patterns indicate a traffic anomaly over a time period, and the causes of these anomalies are worth exploring in the future. Also, it is worth mentioning that we only include 5 categories of traffic-related data. It takes less than one minute for computing regional spatial traffic flow patterns over a 15-minute interval on a PC with intel i7 quad-core 3.7GHZ CPU and 32G RAM. We can improve the accuracy of the identification results of both spatial and temporal pattern identification by including more data categories such as link speed, queue length, etc. If so, the corresponding computational speed may be influenced and one can refer to Akoglu et al. (2012) for fast computation.

Acknowledgement
This research was supported by the Region 2 University Transportation Research Center under project number 49997-32-25.

References


