December 2018

A Framework to Capture Dynamic Traffic Trends from Historical Sensor Data

Carlos Gaviria
espaciocarlos@gmail.com

A FRAMEWORK TO CAPTURE DYNAMIC TRAFFIC TRENDS FROM HISTORICAL SENSOR DATA

By

Carlos Gaviria

Bachelor in Computer Science
College of Engineering
Universidad del Cauca, Colombia
2015

A thesis submitted in partial fulfillment
of the requirements for the

Master of Science in Engineering – Civil and Environmental Engineering

Department of Civil and Environmental Engineering and Construction
Howard R. Hughes College of Engineering
The Graduate College

University of Nevada, Las Vegas
December 2018

UNLV | GRADUATE COLLEGE

Thesis Approval

The Graduate College
The University of Nevada, Las Vegas

November 13, 2018

This thesis prepared by

Carlos Gaviria

entitled

A Framework to Capture Dynamic Traffic Trends from Historical Sensor Data

is approved in partial fulfillment of the requirements for the degree of

Master of Science in Engineering – Civil and Environmental Engineering
Department of Civil and Environmental Engineering and Construction

Alexander Paz, Ph.D.
Examination Committee Chair

David James, Ph.D.
Examination Committee Member

Justin Zhan, Ph.D.
Examination Committee Member

Mohamed Kaseko, Ph.D.
Graduate College Faculty Representative

Kathryn Hausbeck Korgan, Ph.D.
Graduate College Interim Dean

ABSTRACT

The typical approach of increasing infrastructure to alleviate traffic issues such as congestion is becoming unviable due to limited space, high cost, and associated externalities. Control and management strategies using Intelligent Transportation Systems (ITS) seek to maximize the use of existent infrastructure. Many ITS strategies, such as the deployment of information, require or benefit from knowledge about traffic patterns and trends. This study proposes a mathematical programming formulation and solution algorithm that considers multiple time intervals for the estimation of network-wide traffic states and calculation of the corresponding transition probabilities. The proposed solution enables the determination of sections of the network with high traffic variability. This enables the location of congested zones and the determination of reliable traffic flow characteristics. Results from analyzing network level data suggest a trend for congested periods and predominant traffic states in the time intervals considered. It is observed that limited route choices during these periods affected the number of traffic states. From the results set a forecasting system that considers the traffic conditions of multiple time intervals simultaneously was developed and validated with the 10-fold cross validation method. This system presents one of the applications of the joint analysis of clustering of traffic characteristics and associated transition probabilities.

ACKNOWLEDGEMENTS

I am grateful to all of those who mentored and encouraged this and other related research projects during the last years. In particular, I want to express my deep gratitude to Dr. Paz who guided me during my master's degree process. I also want to thank the members of my committee- Dr. Mohamed Kaseko, Dr. Dave James, and Dr. Justin Zhan- for being committed to teaching excellence. I want to thank my friends and colleagues Cristian Arteaga and Victor Molano for their unconditional support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

High costs and unavailability of space constitute important factors that limit increase of highway infrastructure (Padiath, Vanajakshi, & Subramanian, 2012;Ma, Zhou, & Abdulhai, 2015;Hashemi & Abdelghany, 2015). To illustrate a better use of existing infrastructure, various studies have used traffic data to better understand traffic dynamics and generate effective traffic control and management strategies under the umbrella of Intelligent Transportation Systems (ITS) (Paz & Peeta, 2009a; Dimitrakopoulos & Demestichas, 2010; Zavin, Sharif, Ibnat, Abdullah, & Islam, 2017; Dogru & Subasi, 2015). Many ITS strategies require dynamic and network-wide traffic flow data (Paz & Peeta, 2009a; Kachroo, Shlayan, Paz, Sastry, & Patel, 2015; Paz & Peeta, 2009b). Traffic patterns, including variable behavior, can be estimated using historical data to generate effective ITS strategies (Vlahogianni, Karlaftis, & Golias, 2014; Nemtanu, Costea, Badescu, Iordache, & Schlingensiepen, 2016; Van Lint & Hoogendoorn, 2010; Koroliuk & Connaughton, 2015; Paz & Chiu, 2011). In general, traffic data can be classified into location-based, spatial, or network categories (Padiath et al., 2012; Gu, 2016).

There is a body of literature about the analysis and use of this type of traffic data. (Zhang, Ye, Wang, & Malekian, 2016) proposed an algorithm named, grey relational membership degree rank clustering to cluster values of traffic flow characteristics, velocity, density and volume, using sensor data from Nanjing China. The proposed algorithm judged if a congested traffic condition was present by comparing clustered traffic characteristics with those present during a typical congested traffic event. The study obtained better results than previous studies using alternative clustering methods and speed as the only considered traffic characteristic. (Bharadwaj, Biswas, & Ramakrishnan, 2016) created a traffic dataset using video from various locations of a city in India. Samples of images of vehicles were classified with different clustering and artificial intelligence algorithms. The study concluded that further development to obtain more robust classification algorithms was required, as the various tested clustering methods provided low accuracy results. (Dogru & Subasi, 2015) Generated traffic data of speed and location of vehicles from a simulation model. The data, were analyzed with the use of various clustering methods to determine anomalous situations. If the anomalies were present during a prolonged period of time, the system defined the presence of an accident on the highway. The study concluded that the best clustering methods to determine the presence of accidents on a highway are the “density-based spatial clustering of applications with noise (DBSCAN)” and the “agglomerative hierarchical clustering (AHC)”. These studies showed how aggregation of traffic data contributes to acquire knowledge or information regarding traffic conditions. Nonetheless, these studies provide a limited view because frequency of traffic conditions is not considered. Analyzing this frequency would result in the estimation of new patterns and knowledge that can determine how common are certain traffic conditions in a corridor or network.

The existing literature provides methods to forecast traffic states and generate traffic control and management strategies (Zhao, 2015; J. Xu, Deng, Demiryurek, Shahabi, & Schaar, 2015; Allström et al., 2016; Oh, Byon, & Yeo, 2016; Barros, Araujo, & Rossetti, 2015). An important group of studies involving traffic guidance require effective forecasting systems (Zavin et al., 2017; Paz & Peeta, 2009d; Paz & Peeta, 2009c; Paz & Peeta, 2008; Lu, Duan, & Zheng, 2009; Boriboonsomsin, Barth, Zhu, & Vu, 2012). Using historical traffic data, a knowledge base can be built and used in conjunction with real-time traffic data to forecast and advise drivers about traffic conditions and adequate traffic routes. A system framework capable of determining possible traffic states across time intervals can be used to calculate transition probabilities where the highest values denote likely future conditions. Transition probabilities among traffic states can be used to estimate the probability of having a specific group of traffic characteristics at a specific time period (Gu et al., 2016).

(Y. Xu, Xi, & Li, 2016) developed a scheme of traffic signals using a transition probability model applied to a network of four segments. The transition probability model was linked with a Markov Decision Process to control traffic signals in the network. The transition probability model, described with a mathematical programming formulation, was first applied to a link and extended to the four segments of the network. To test this scheme, simulation experiments were performed. Results illustrated that this scheme outperformed previous related methods. (DOU, WANG, & GUO, 2011) developed a traffic guidance system to support travelers and managers of a traffic network. Logistic regression was used to build a probability function to estimate traffic state transitions. A historical traffic dataset was used to experiment with the estimated transitions in time intervals of 5, 10, 15 and 30 min. The study provided better results while estimating traffic state transitions using time intervals of 5 minutes. (Gu et al., 2016) developed a probabilistic model to estimate traffic states. This model generated a distribution using the Gibbs sampling method. This distribution was generated from previously observed traffic states. The process of estimating future traffic states started with the determination of the current conditions; then, a probability distribution used the current traffic conditions as an input to determine a future traffic state. Various experiments tested the efficiency of the model. (Hellinga & Noroozi, 2014) developed a near-future predictor using a Markov model to estimate transition probabilities among different traffic states. Different strategies were proposed to decrease error. The proposed approach was tested through experimentation with highway traffic data. The results indicated that the approach was capable of predicting traffic states and that the postprocessing strategies to decrease the error in the predictor were successful.

The existing literature has proposed various models that used the calculation of transition probabilities among traffic states to control and manage traffic networks (Zhao, 2015; J. Xu et al., 2015; Allström et al., 2016; Oh et al., 2016; Barros et al., 2015; Maheshwari, Kachroo, Paz, & Khaddar, 2015). Common limitations of these models are related to the small size of the network considered. Further, the transition probabilities were calculated between a specific time interval *t* and its subsequent *t+1*; however, subsequent time intervals, *t+2, t+3,…,t+n* were not considered. This presents a limited view of what could be the evolution of traffic states. The nonexistence of a joint analysis of clusters of traffic characteristics and associated transition probabilities limits the understanding of existent traffic trends in historical traffic datasets. Important trends such as congested periods, predominant traffic states, and transition probabilities could be obtained from this analysis and used to develop improved control and management systems such as, traffic guidance systems, traffic signal systems, and forecasting systems (core components of the ITSs) that can be used by drivers and traffic controllers to take decisions, avoid unwanted traffic conditions, or even lead the system to a desired traffic state. The development of these systems can be improved considering the possible evolution of traffic conditions. This would offer more accurate results than the ones obtained by actual systems that only factor for traffic conditions at a given time and a subsequent. Not considering the evolution of traffic condition between time intervals creates a gap for uncertainty that could lead to inconsistent results, this presents a problem for actual control and management systems.

To address the described limitations, the present work proposes a mathematical programming formulation and solution algorithm that considers multiple time intervals for the generation of clustered traffic characteristics and the calculation of transition probabilities among them. Centroids of the clustered traffic characteristics represent average traffic states for each time interval. A forecasting system based in the obtained results is presented.

The remaining portions of this manuscript is divided into four sections. Chapter 2 presents the methodology section, where a clustering problem is defined using a mathematical programming formulation and a proposed calculation for transition probabilities is described. Chapter 3 describes a solution algorithm for the proposed mathematical program. Chapter 4 includes numerical experiments and results using data from the freeway of southern Nevada, Nevada. Results obtained are used to locate sections with high variations of traffic characteristics and to generate a forecasting system. The final part of this chapter contains a discussion of the obtained results. Chapter 5 provides conclusions, future work and limitations of the proposed framework.

CHAPTER 2: METHODOLOGY

Terms used in the following mathematical programming formulation are defined in Table 1. The following is a discrete time non-linear mixed integer mathematical programming formulation. Superscript t, represents a time interval with a length Δ which is considered as the period used to register and average observations of various traffic characteristics in the dataset. It is assumed that the length of the observation time Δ is the same for every traffic characteristic. The problem of the generation of clusters of traffic characteristics and its associated transition probabilities is described as follows. First a definition of the terms is displayed. Second a description of the mathematical programming formulation for the generation of cluster of traffic characteristics is presented and notated in equations. Finally, the problem of the computation of transition probabilities between traffic states is described by an algorithm of five steps; the solution of the described problem is presented in the chapter 3.

Table 1. Definition of Terms

Term	Definition
N	Set of nodes in the network
R	Set of upstream nodes, $R \subseteq N$
r	Superscript to denote an upstream node, $r \in R$
P	Set of downstream nodes, $P \subseteq N$
p	Superscript to denote a downstream node, $p \in P$
A	Set of links in the network
rp	Superscript to denote a link between nodes r and p, $rp \in A$
Y	Set of years considered in the analysis
y	Superscript to denote a year, $y \in Y$
S	Set of seasons in year y
s	Superscript to denote a season in a year, $s \in S; s = 1,2,3,4$
W	Set of weeks in season s
w	Superscript to denote a week in a season, $w \in W;\ w = 1,\ldots\ldots\ldots,12$
D	Set of days in week w
d	Superscript to denote a day in a week, $d \in D;\ d = 1,\ldots\ldots\ldots,7$

T	Time period of analysis
t	Superscript to denote an observation time interval, $t \in T$; $t = 1,\ldots\ldots,T$
Δ	Length of an observation time interval t
M	Historic network-wide traffic flow characteristics including speed, volume, and occupancy.
c	Subscript to denote a traffic characteristic such as speed from M
$m_c^{yswdt,rp}$	Traffic characteristic c for link rp at $yswdt$, $m_c^{yswdt,rp} \in\ M$
m^{yswdt}	Network-wide traffic characteristics $m_c^{yswdt,rp}$ at $yswdt$, $\forall\ c \in M, \forall\ r \in R, \forall\ p \in P$
B^{ydt}	Set of Network-wide traffic characteristics m^{yswdt} at ydt, where: $B^{yt} = \bigcup_{s=1}^{S} \bigcup_{w=1}^{W} m^{yswdt}$ $\forall\ y \in Y, d \in D, \forall\ t \in T$
Q^{ydt}	Number of clusters of traffic characteristics in $B^{ydt}, \forall\ y \in Y, \forall\ d \in D, \forall\ t \in T$
$q(y,d,t)$	Subscript to denote a cluster of network-wide traffic flow characteristics in B^{ydt}, $q(y,d,t) = 1,\ldots\ldots,Q^{ydt}\ ; \forall\ y \in Y, \forall\ d \in D, \forall\ t \in T$
$K^{ydt}_{q(y,d,t)}$	Cluster q of network-wide traffic flow characteristics at ydt, where: $(m^{yswdt} \in K^{ydt}_{q(y,d,t)}) \Leftrightarrow (\delta^{yswdt}_{q(y,d,t)} = 1)$ $\forall\ y \in Y, \forall\ s \in S, \forall\ w \in W, \forall\ d \in D, \forall\ t \in T, \forall q(y,d,t) = 1 \ldots Q^{ydt}$
K^{ydt}	Set of clusters of network-wide traffic flow characteristics at ydt
$a^{ydt}_{q(y,d,t)}$	Average traffic state for cluster $K^{ydt}_{q(y,d,t)}$ defined as its centroid, $\forall q(y,d,t) = 1 \ldots Q^{ydt}, \forall\ y \in Y, \forall\ d \in D, \forall\ t \in T$
$\delta^{yswdt}_{q(y,d,t)}$	Indicator variable: $\delta^{yswdt}_{q(y,d,t)} \begin{cases} 1 \Leftrightarrow \left(m^{yswdt} \in K^{ydt}_{q(y,d,t)}\right) \\ 0 \quad otherwise \end{cases}$ $\forall\ y \in Y, \forall\ s \in S, \forall\ w \in W, \forall\ d \in D, \forall\ t \in T, \forall q(y,d,t) = 1 \ldots Q^{ydt}$
$\rho_{q(y,d,t)\ \ q'(y,d,t+1)}$	Notation for the calculation of transition probability from any cluster $q(y,d,t)$ in K^{ydt} to any cluster $q'(y,d,t+1)$ in $K^{yd(t+1)}$

Mathematical Programming - Clustering

The objective function to generate clusters K^{ydt}:

$$\min \sum_{s}^{S} \sum_{w}^{W} \sum_{d}^{D} \sum_{q(y,d,t)}^{Q^{ydt}} (m^{yswdt} - a_{q(y,d,t)}^{ydt})^2 \cdot \delta_{q(y,d,t)}^{yswdt} \quad \forall\, y \in Y, \forall\, t \in T \tag{1}$$

Subject to the following:
Definitional Constraint:

$$a_{q(y,d,t)}^{ydt} = \frac{1}{\sum_s \sum_w \delta_{q(y,d,t)}^{yswdt}} \cdot \sum_{r}^{R} \sum_{p}^{P} \sum_{s}^{S} \sum_{w}^{W} \sum_{c}^{C} (m_c^{yswdt,rp} \cdot \delta_{q(y,d,t)}^{yswdt}) \quad \begin{matrix} \forall\, y \in Y, \forall\, d \in D, \\ \forall\, t \in T \end{matrix} \tag{2}$$

$$\delta_{q(y,d,t)}^{yswdt} \begin{cases} 1 \Leftrightarrow \left(m^{yswdt} \in K_{q(y,d,t)}^{ydt}\right) \\ 0 \quad otherwise \end{cases} \quad \begin{matrix} \forall q(y,d,t) = 1 \ldots Q^{ydt}, \forall\, y \in Y, \forall\, s \in S, \\ \forall\, w \in W, \forall\, d \in D, \forall\, t \in T \end{matrix} \tag{3}$$

$$\sum_{q(y,d,t)}^{Q^{ydt}} \delta_{q(y,d,t)}^{yswdt} = 1 \quad \forall\, y \in Y, \forall\, s \in S, \forall\, w \in W, \forall\, d \in D, \forall\, t \in T \tag{4}$$

$$\sum_{s}^{S} \sum_{w}^{W} \sum_{d}^{D} \delta_{q(y,d,t)}^{yswdt} > 0 \quad \forall\, y \in Y, \forall\, t \in T \tag{5}$$

The Equation 1 describes a minimization process between each m^{yswdt} and its centroid $a_{q(y,d,t)}^{ydt}$. This minimization enables the decision variable $\delta_{q(y,d,t)}^{yswdt}$ to indicate whether m^{yswdt} is part of cluster $K_{q(y,d,t)}^{ydt}$. The Equation 2 is used as a definitional constraint to generate centroid $a_{q(y,d,t)}^{ydt}$ for cluster $K_{q(y,d,t)}^{ydt}$, each centroid contains the averaged results of the m^{yswdt} that belongs to cluster $K_{q(y,d,t)}^{ydt}$. The divisor of this equation serves as a counter of how many traffic characteristics of an specific type exist in a cluster, using the dividend part of the equation, the average of the different traffic characteristics are used for the creation of traffic states. The Equation 3 is a constraint that determines whether m^{yswdt} is part of the cluster $K_{q(y,d,t)}^{ydt}$. The

Equation 4 is a constraint used to ensure that each m^{yswdt} is a member of only one cluster $K^{ydt}_{q(y,d,t)}$. The Equation 5 is a constraint that ensures that at least one m^{yswdt} is part of a cluster $K^{ydt}_{q(y,d,t)}$.

Clustering Problem

Given a vector of network-wide traffic flow characteristics m^{yswdt}, the problem is to determine clusters of data that summarize time-dependent average network states denoted by terms $\delta^{yswdt}_{q(y,d,t)}, a^{ydt}_{q(y,d,t)}, Q^{ydt}, and\ K^{ydt}$. Clusters K^{ydt} composed by network-wide traffic flow characteristics m^{yswdt} generates centroids $a^{ydt}_{q(y,d,t)}$ representing averaged network states from the traffic characteristics. The number of network-wide traffic flow characteristics considered for clusters K^{ydt} can be calculated as $|S| * |W| * |D|$. Therefore, it is assumed that traffic flow characteristics for the same day and time during a year are likely to be similar. Indicator $\delta^{yswdt}_{q(y,d,t)}$ defines which network-wide traffic flow characteristics m^{yswdt} are part of clusters K^{ydt}. The definition of the number of clusters Q^{ydt} is a fundamental issue for clustering processes. This study assesses the most convenient number of clusters for set B^{ydt} with the silhouette method (Rousseeuw, 1987). Details about the selection and usage of this method are provided in the solution section.

Transition Probabilities Problem

Given clusters K^{ydt} and $K^{yd(t+1)}$, the problem is to calculate transition probabilities $\rho_{q(y,d,t)\ \underset{-}{q'}(y,d,t+1)}$. K^{ydt} and $K^{yd(t+1)}$ include m^{yswdt} and $m^{yswd(t+1)}$ respectively. Transition probabilities are calculated counting the number of occurrences of $m^{yswdt} \in K^{ydt}_{q(y,d,t)}$ and $m^{yswd(t+1)} \in K^{yd(t+1)}_{q(y,d,t)}$.

CHAPTER 3: PROPOSED SOLUTION

Clustering Generation

To select an appropriate solution for the clustering problem different approaches were evaluated. The first attempt to solve this problem tried to cluster traffic characteristics m^{yswdt} based on prespecified ranges for the traffic characteristics for all the links in the network. However, the values of the traffic characteristics were falling outside the prespecified ranges in 99% of the cases, making this solution unviable. From this point, different clustering methods were tested and evaluated using the silhouette width method which is a widely used goodness-of-fit measure that can be used with multiple purposes in clustering problems. For example, selection of clustering algorithms or assessing of number of clusters in a dataset. The different average silhouette widths obtained from processing a sample of multiple entries of the dataset are displayed in the Table 2.

Table 2. Clustering Methods Silhouette Width

Clustering Method	**Average Silhouette Width**
hierarchical clustering	0.10
Enhanced hierarchical clustering	0.11
k-medoids/pam clustering	0.13
k-means	0.32

Hard clustering methods enable the formation of clusters with members that belong to a unique cluster. Given the nature of the problem here stablished, only hard clustering methods were considered. The DBSCAN clustering method was not considered given its inability to use multiple processors simultaneously, the size of the dataset to be processed makes of it an unviable solution (Ahmad & Dang, 2015). The K-Means method presented the best silhouette width among the different methods tested, because of this it is the method selected to solve the mathematical programming formulation.

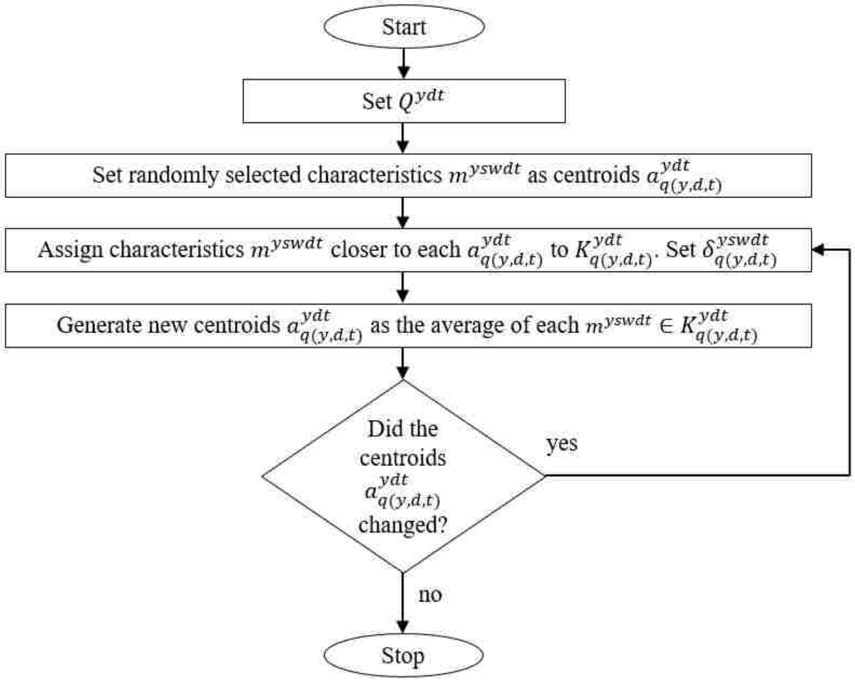
Values for $\delta_{q(y,d,t)}^{yswdt}$, $a_{q(y,d,t)}^{ydt}$, Q^{ydt} and K^{ydt} need to be determined. A standard K-means clustering algorithm that has been used to cluster traffic data in previous studies (Nawrin, 2017; Montazeri-Gh & Fotouhi, 2011) was used. The first step to generate clusters with the K-means algorithm is to set the value for the number of clusters Q^{ydt}. The silhouette width method, is used as a goodness-of-fit measure that enables the assessing of the optimal number of clusters. Silhouette width (Nawrin, 2017; Yingqiu, Wei, & Yunchun, 2007) should be calculated for various numbers of clusters with the same data. The highest width obtained during the search determines the optimal number of clusters. The silhouette width method uses simultaneously separation and cohesion. To better understand its standard computation, a three steps algorithm is described as follows. *1)* The average distance from an element *i* th to every other element in its cluster is calculated and assigned to $a_{i.}$ *2)* For each cluster that does not contain the *i* th element, calculate the average distance from the *i* th element to all the elements of each cluster. Assign b_i as the minimum average distance found. *3)* The silhouette width coefficient for the element *i* th is given by: s_i = (b_{i} - a_i) / max($a_{i,}$ b_i). The silhouette width value has a range of -1 to 1, it is desirable for its value to be positive, meaning that the average distance (a_i) from an element *i* th to every other element in its cluster, is smaller than the minimum average distance (b_i) to elements of other clusters. It is also desirable for a_i to be near 0, this because the silhouette width assumes a value of 1, its maximum, with a_i = 0.

Once the number of clusters Q^{ydt} is determined, the values for terms $\delta_{q(y,d,t)}^{yswdt}$, $a_{q(y,d,t)}^{ydt}$, and K^{ydt} can be searched with the K-means algorithm. The flowchart depicted in Figure 1, displays a series of steps followed by the algorithm to form clusters K^{ydt} and set the values for $\delta_{q(y,d,t)}^{yswdt}$ and $a_{q(y,d,t)}^{ydt}$. To start, the algorithm requires the definition of an initial set of centroids $a_{q(y,d,t)}^{ydt}$ for each $K_{q(y,d,t)}^{ydt}$ which are formed by randomly selected network-wide traffic flow characteristics m^{yswdt}. A loop with the following three steps is repeated until convergence is achieved when centroids stop changing across iterations with a tolerance of 0.00001%, or until a maximum of 20 iterations is achieved.

1. Characteristics m^{yswdt} that are closer to centroids $a_{q(y,d,t)}^{ydt}$ are set into clusters $K_{q(y,d,t)}^{ydt}$. If $m^{yswdt} \in K_{q(y,d,t)}^{ydt}$, then $\delta_{q(y,d,t)}^{yswdt} = 1$; otherwise, $\delta_{q(y,d,t)}^{yswdt} = 0$.
2. An average of characteristics m^{yswdt} that belong to cluster $K_{q(y,d,t)}^{ydt}$ is defined as the new centroid $a_{q(y,d,t)}^{ydt}$ of the cluster.
3. Check if centroid $a_{q(y,d,t)}^{ydt}$ of the cluster changed; if it did not, stop.

Once the algorithm finishes, the corresponding values for $\delta_{q(y,d,t)}^{yswdt}$, $a_{q(y,d,t)}^{ydt}, and\ K^{ydt}$ are assigned. The silhouette method can be applied to measure the goodness-of-fit of the results of the clustering process.

Figure 1. K-means applied to cluster characteristics m^{yswdt}


Start
Set Q^{ydt}
Set randomly selected characteristics m^{yswdt} as centroids $a^{ydt}_{q(y,d,t)}$
Assign characteristics m^{yswdt} closer to each $a^{ydt}_{q(y,d,t)}$ to $K^{ydt}_{q(y,d,t)}$. Set $\delta^{yswdt}_{q(y,d,t)}$
Generate new centroids $a^{ydt}_{q(y,d,t)}$ as the average of each $m^{yswdt} \in K^{ydt}_{q(y,d,t)}$
Did the centroids $a^{ydt}_{q(y,d,t)}$ changed?
yes
no
Stop

Transition Probabilities Algorithm

An algorithm for the calculation of transition probabilities between the clusters of K^{ydt} and $K^{yd(t+1)}$ is implemented as follows:

1. Set values for the year *y*, the day *d*, and the time intervals *t* and *t+1*.
2. Set *e* as the number of occurrences where the indicators variables

   $\delta^{yswdt}_{q(y,d,t)} = 1$ and $\delta^{yswd(t+1)}_{q\prime(y,d,t)} = 1$

   $\forall\ y \in Y, \forall\ s \in S, \forall\ w \in W, \forall\ d \in D, \forall q = 1 \ldots Q^{ydt}, \forall q' = 1 \ldots Q'^{yd(t+1)}$
3. Set *h* as the total number of days considered.
4. The transition probability between *t* and *t+1* is calculated as *e / h*.
5. Repeat steps 1 to 4 for each $t \in T$.

The transition probability is calculated between the sets of clusters K^{ydt} and $K^{yd(t+1)}$. Step one sets the required values for the superscripts of these sets. The second step sets e as a counter of occurrences where Indicators $\delta_{q(y,d,t)}^{yswdt} = 1$ and $\delta_{q\prime(y,d,t)}^{yswd(t+1)} = 1$. This happens when $m^{yswdt} \in K_{q(y,d,t)}^{ydt}$ and $m^{yswd(t+1)} \in K_{q(y,d,t)}^{yd(t+1)}$, respectively. The third step sets h as the total number of days considered, that is, the sample of size $|S| * |W| * |D|$. The fourth step calculates the transition probability between $K_{q(y,d,t)}^{ydt}$ to $K_{q\prime(y,d,t)}^{yd(t+1)}$ as e / h. Finally, step 5 repeats steps 1 to 4 for each time interval considered. A zero-probability value means that there were not occurrences between the clusters.

CHAPTER 4: EXPERIMENTS AND RESULTS

Specifications of software and hardware used to execute experiments

Software:

- Operating system: CentosOS 7.5
- Programming Language: R 3.5.1, algorithms written with multiprocessing capabilities.
- Data Base Management System: Microsoft Access 2016 (21.5 Gb)
- File system (data queried from DBMS): csv files

Hardware:

- Number of CPUs: 8
- CPUs Used: Intel(R) Xeon(R) CPU E7- 4870 @ 2.40GHz (10 cores)
- Total cores: 80 cores
- RAM Memory: 256 Gb DDR3 1333 MHz

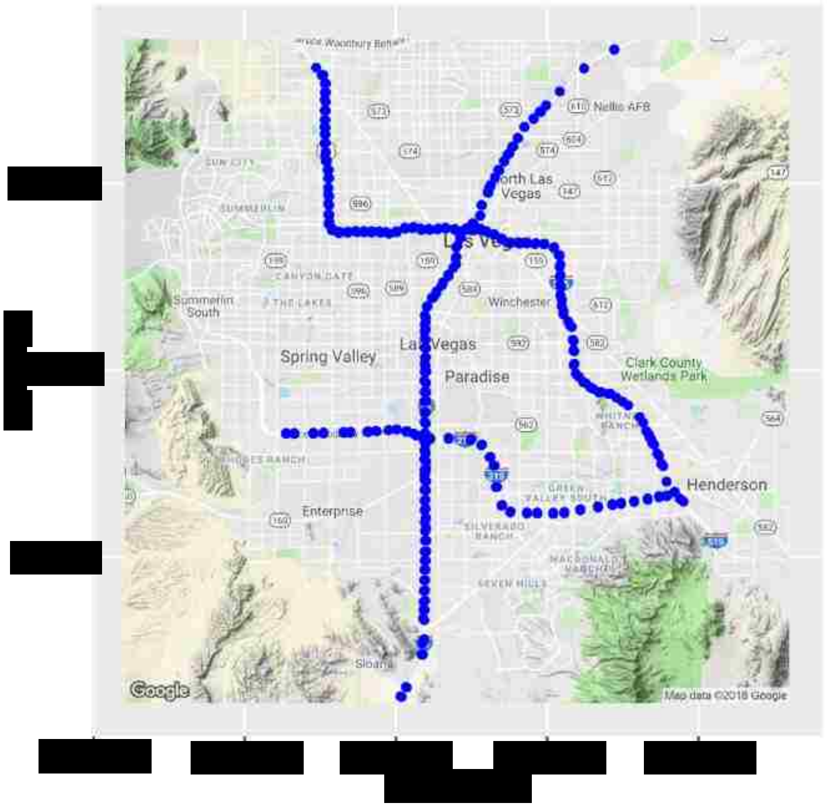
The approximated time to compute the different clusters, transition probabilities and validation processes in the time intervals considered was 6 days (144 hours). The processes that demanded 90% of the processing time were related to partition of data, processes like this are sequential and cannot be easily parallelized to take advantage of the multiple processors. Other processes such as querying the database for the generation of a csv file that contains the final dataset are not factored in that 144 hours, the approximated time of processing these other calculations was 10 hours giving a total of 154 hours of continuous processing in the detailed hardware and software.

Data

The Freeway & Arterial System of Transportation (FAST, NV, 2018) provided the traffic dataset used in this study. The main characteristics of this traffic dataset are:

- The dataset contains information collected from September of 2016 to August of 2017.
- The dataset includes information of 466 sensors located across the freeway system in Las Vegas Nevada. The precise location of these sensors is determined by latitude and longitude coordinates. Figure 2 provides a map view representing the location of each sensors with a blue marker. The spacing between many of these sensors is approximately 0.33 miles.
- The dataset contains average values of speed, volume, and occupancy. Each value is the result of averaging information collected in time intervals of 15 minutes from each lane where the sensor is located.
- Speed is measured in miles per hour; volume is reported as the total number of vehicles during a time interval, and the occupancy is the seconds during which a sensor is occupied by a vehicle.
- The size of the dataset is 21.5 GB, containing an approximated total of 16,328,640 records. Each record includes sensor Id, timestamp, speed, volume, and occupancy.
- Sensor dataset incorporates freeways I-15, I-515, US-95, U-215

Figure 2. Location of sensors that provided the data for the dataset



[redacted] [redacted] [redacted] [redacted] [redacted] [redacted] [redacted]

Experimental Setup

Data from Mondays to Thursdays were selected for the analyses in this study and set as d. It is expected that during these days different traffic characteristics will reflect common commuting patterns. A total of 16 time intervals t starting at 15:00 and ending at 19:00 were considered for the generation of clusters $K_{q(y,d,t)}^{ydt}$ of network-wide traffic characteristics m^{yswdt}. These specific time intervals were selected because they cover the transition from a non-peak hour to a peak hour, enabling the generation of results with singular characteristics. Given these considerations and the provided information of the dataset, the values for subscripts y, s, w, d, t and c were set as follows:

$y = 1,2$	Years 2016 and 2017, respectively
$s = 3,4,1,2$	Seasons Fall, Winter, Spring, and Summer, respectively
$w = 1, \dots ,12$	Weeks 1 to 12 of each season with three months per season
$d = 1,2,3,4$	1 = Monday, 2 = Tuesday, 3 = Wednesday, and 4 = Thursday
$t = 61, \dots ,77$	61 = 14:45 to 15:00,……, 77 = 18:45 to 19:00
$c = 1,2,3$	Speed, Volume, and Occupancy, respectively

The dates that determine each change of season were slightly modified to fit with the proposed layout for the data with 12 weeks per season:

$s = 1$	Spring	13 March to 4 June, 2017
$s = 2$	Summer	5 June to 27 August, 2017
$s = 3$	Fall	26 September to 18 December, 2016
$s = 4$	Winter	19 December to 12 March, 2016 – 2017

The total number of days considered for this study was 192 with 4 days per week (Monday thru Thursday) during 48 weeks. The clusters for each time interval t were conformed by the information provided by the traffic characteristics of these 192 days.

Clustering at network level

The proposed solution uses an algorithm with iterative steps that generated a vast number of results. Two samples of results were selected to be described in detail. The first sample corresponds to results obtained processing data for time interval $t = 68$ (16:30 to 16:45), where 6 clusters were generated, allowing the observation of different traffic states. The second sample corresponds to results obtained processing the data for time interval $t = 74$ (18:00 to 18:15). This time interval is part of the time intervals that are between $t = 69$ (16:45 to 17:00) and $t = 75$ (18:15 to 18:30) where only 2 clusters were generated for each time interval.

Results for the first sample $t = 68$ (16:30 to 16:45)

Figure 3 displays the value of the total average silhouette obtained after applying K-means with the number of clusters Q^{ydt} varying from 2 to 10. The highest average silhouette width corresponds to $Q^{ydt} = 6$. Hence, this is the number of clusters to be generated for this time interval. Figure 4 describes clusters K^{ydt} obtained after applying the K-means algorithm to the 192 network wide traffic characteristics m^{yswdt} with $t = 68$, $y = 2016\ to\ 2017$, and $Q^{ydt} = 6$. Using the silhouette width, each line in the figure represents how well each m^{yswdt} fits within its cluster $K^{ydt}_{q(y,d,t)}$.

Each cluster $K^{ydt}_{q(y,d,t)}$ generated a centroid $a^{ydt}_{q(y,d,t)}$ or average traffic state that contains the mean of traffic characteristics m^{yswdt} that belong to the cluster. A map showing the speed per segment for each measured centroid $a^{ydt}_{q(y,d,t)}$ is depicted in Figure 5. Table 3 provides details about the aggregated traffic characteristics and cluster information for each traffic state.

Figure 3. Number of clusters versus Silhouette width for $t = 68$

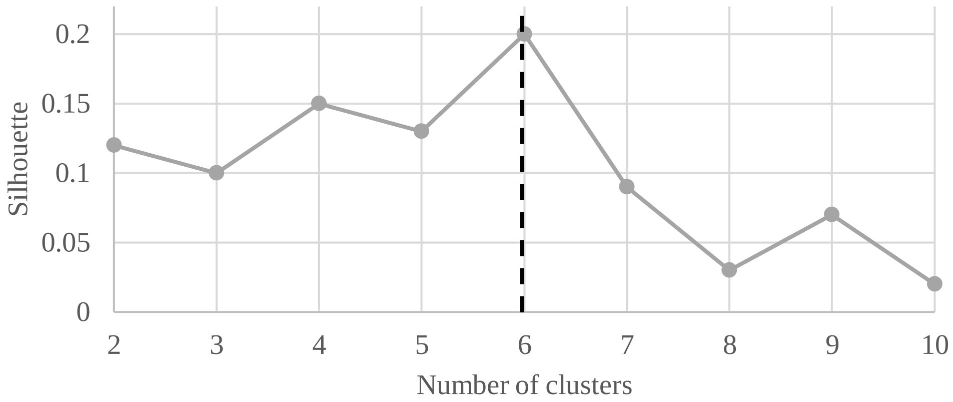
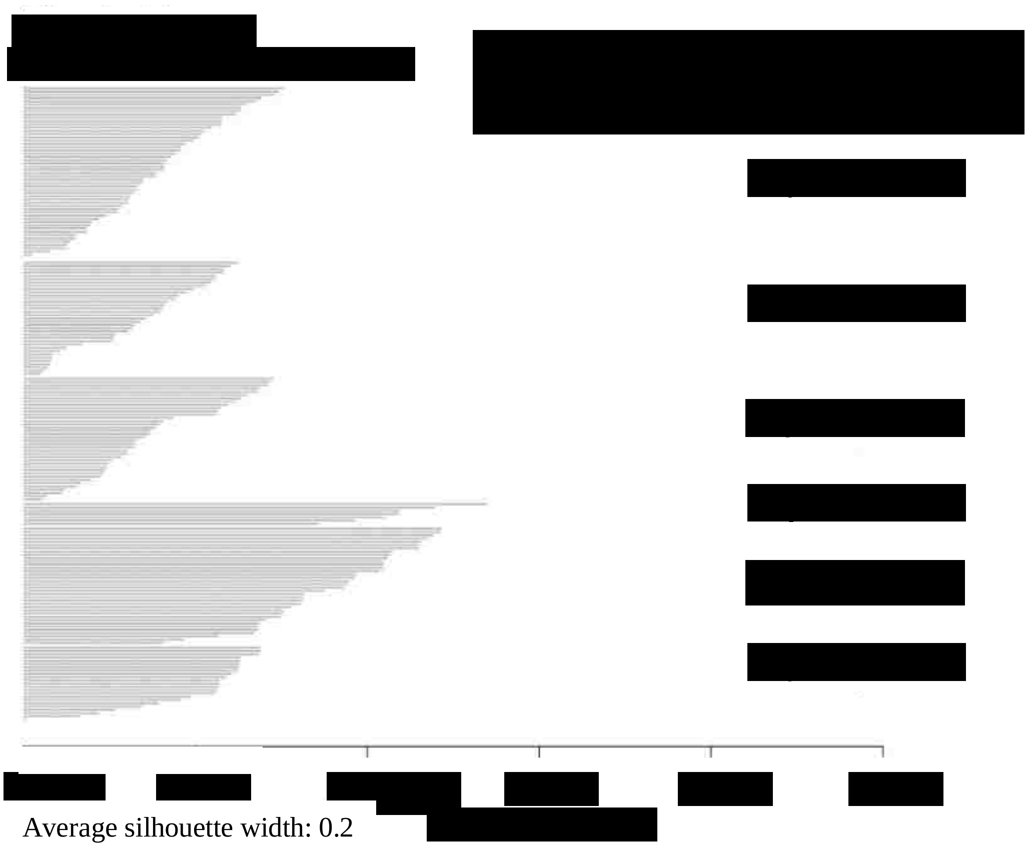
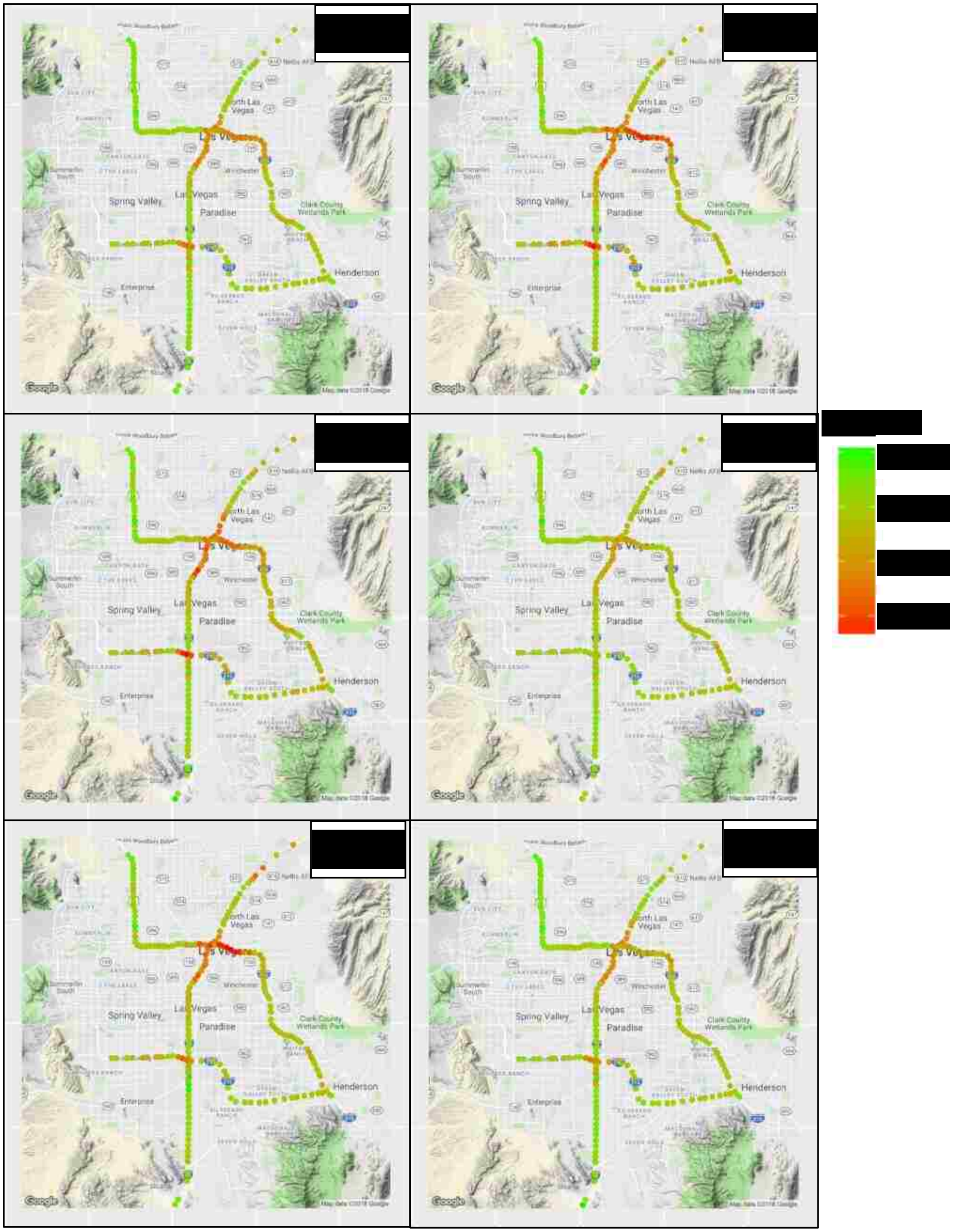


Figure 4. Silhouette obtained for each cluster in the time interval $t = 68$



[redacted] [redacted] [redacted] [redacted] [redacted] [redacted] [redacted] [redacted] [redacted] [redacted] [redacted] [redacted] [redacted] [redacted]

Figure 5. Traffic states $a^{ydt}_{1(y,d,t)}$ - $a^{ydt}_{6(y,d,t)}$ measured for the time interval $t = 68$



[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

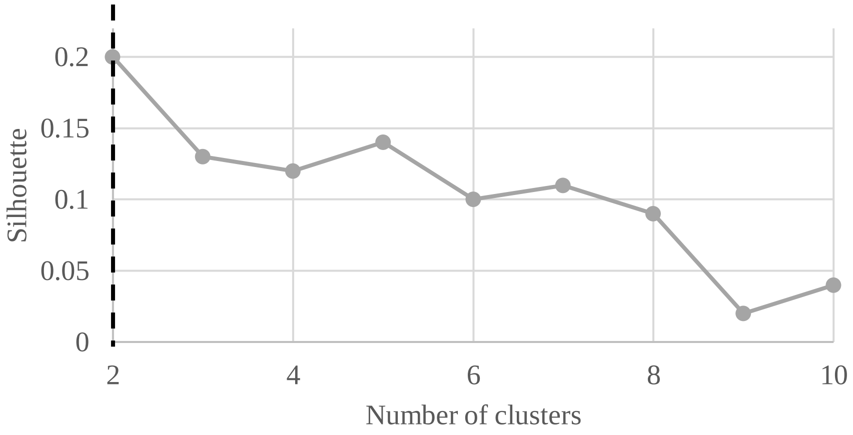
Table 3. Measured traffic characteristics and cluster information for time interval $t = 68$

Traffic state	Cluster	Number of characteristics m^{yswdt} in cluster	Average Speed	Average Volume	Average Occupancy	Silhouette Width
$a_{1(y,d,t)}^{ydt}$	$K_{1(y,d,t)}^{ydt}$	53	57.21 mph	941 vehicles	9.53 seconds	0.16
$a_{2(y,d,t)}^{ydt}$	$K_{2(y,d,t)}^{ydt}$	35	58.65 mph	947 vehicles	9.36 seconds	0.13
$a_{3(y,d,t)}^{ydt}$	$K_{3(y,d,t)}^{ydt}$	38	58.25 mph	926 vehicles	9.42 seconds	0.15
$a_{4(y,d,t)}^{ydt}$	$K_{4(y,d,t)}^{ydt}$	7	64.00 mph	796 vehicles	6.31 seconds	0.3
$a_{5(y,d,t)}^{ydt}$	$K_{5(y,d,t)}^{ydt}$	36	59.19 mph	919 vehicles	8.67 seconds	0.36
$a_{6(y,d,t)}^{ydt}$	$K_{6(y,d,t)}^{ydt}$	23	57.69 mph	965 vehicles	9.48 seconds	0.2

Results for second sample $t = 74$ (18:00 to 18:15)

Figure 6 shows the value of the total average silhouette obtained after applying K-means with the number of clusters Q^{ydt} varying from 2 to 10.

Figure 6. Number of clusters versus Silhouette width for $t = 74$



The highest average silhouette width corresponds to $Q^{ydt} = 2$. Hence, this is the number of clusters to be generated for this time interval. Figure 7 describes clusters K^{ydt} obtained after applying the K-means algorithm to the 192 network wide traffic characteristics m^{yswdt} with $t = 74, y = 2016\ to\ 2017$, and the number of clusters $Q^{ydt} = 2$. Using the silhouette width as a goodness of fit measure, each line in the figure represents how well each m^{yswdt} fits within its cluster $K^{ydt}_{q(y,d,t)}$. A map showing the measured speed per segment for each centroid or traffic state $a^{ydt}_{q(y,d,t)}$ of each cluster $K^{ydt}_{q(y,d,t)}$ is depicted in Figure 8. Table 4 provides details about the aggregated traffic characteristics and cluster information for each traffic state.

Figure 7. Silhouette obtained for each cluster in the time interval $t = 74$

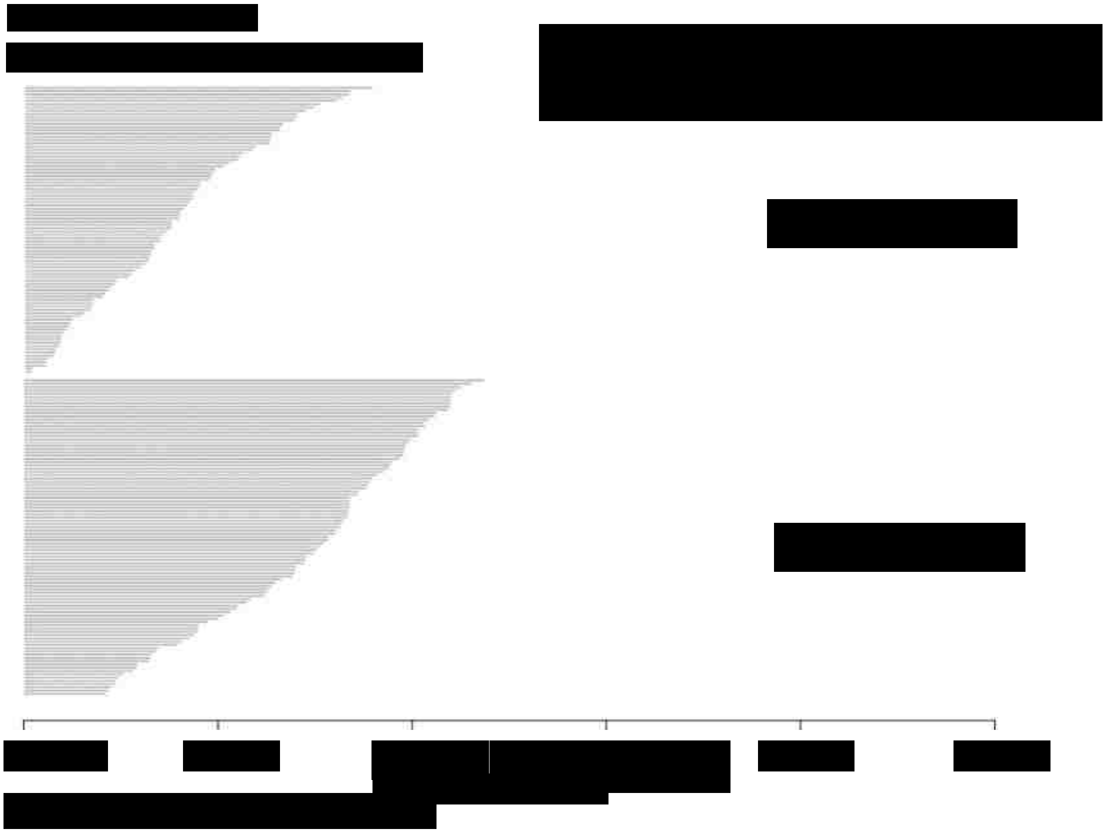


Figure 8. Traffic states a_1^{ydt} and a_2^{ydt} measured in the time interval $t = 74$

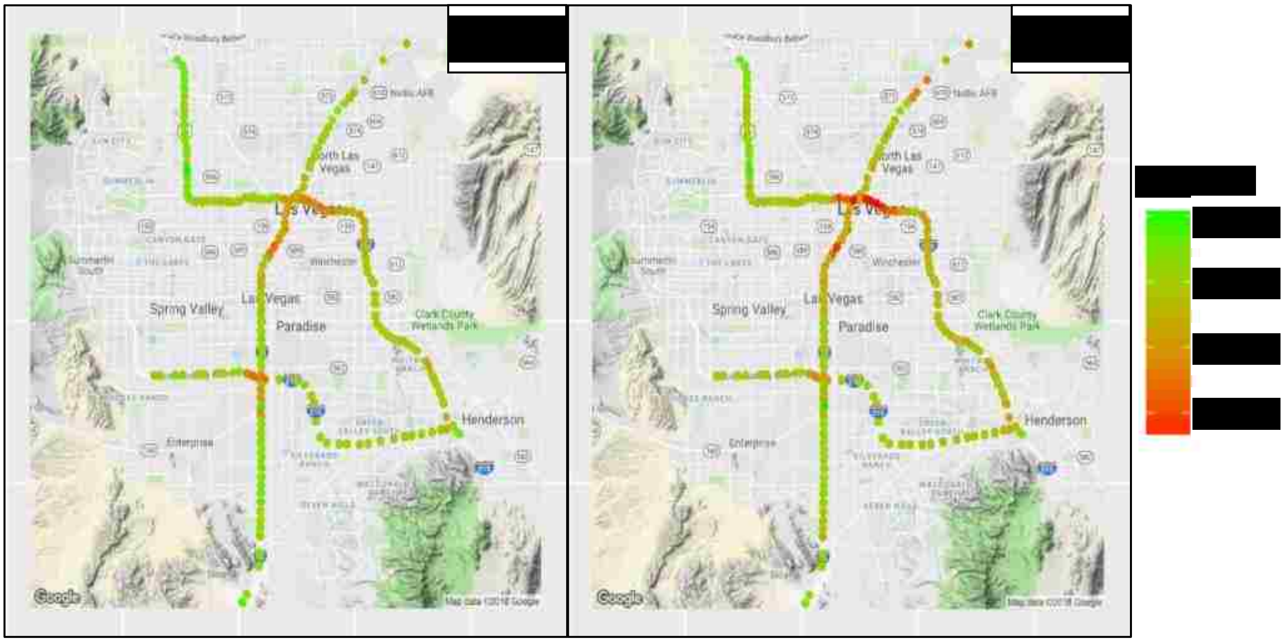


Table 4. Measured traffic characteristics and cluster information for time interval $t = 74$

Traffic state	Cluster	Number of characteristics m^{yswdt} in cluster	Average Speed	Average Volume	Average Occupancy	Silhouette Width
$a_{1(y,d,t)}^{ydt}$	$K_{1(y,d,t)}^{ydt}$	95	60.18 mph	845 vehicles	7.63 seconds	0.13
$a_{2(y,d,t)}^{ydt}$	$K_{2(y,d,t)}^{ydt}$	97	62.76 mph	832 vehicles	6.76 seconds	0.29

Transition probabilities at the network level

After generating clusters K^{ydt} for the time intervals considered, transition probabilities were calculated using the steps of the proposed algorithm. Figures 9 to 12 display a summary of results from the clustering process and the calculated transition probabilities during the corresponding time intervals. In these figures, each circle represents a cluster $K_{q(y,d,t)}^{ydt}$ at time interval depicted on the horizonal bar at the bottom of the figure. The number inside each circle or cluster, represents the percentage of traffic characteristics m^{yswdt} that is included in that cluster with respect of the others for the same time interval. The dashed lines and its percentage values are the calculated transition probabilities from each cluster in t to any other cluster in $t+1$. The solid line, represents the highest transition probability between each time interval. Transition probabilities equal to 0% are not displayed in the figures.

Figure 9. Transition probabilities across time intervals $\boldsymbol{t = 15{:}00}$ to $\boldsymbol{t = 16{:}00}$

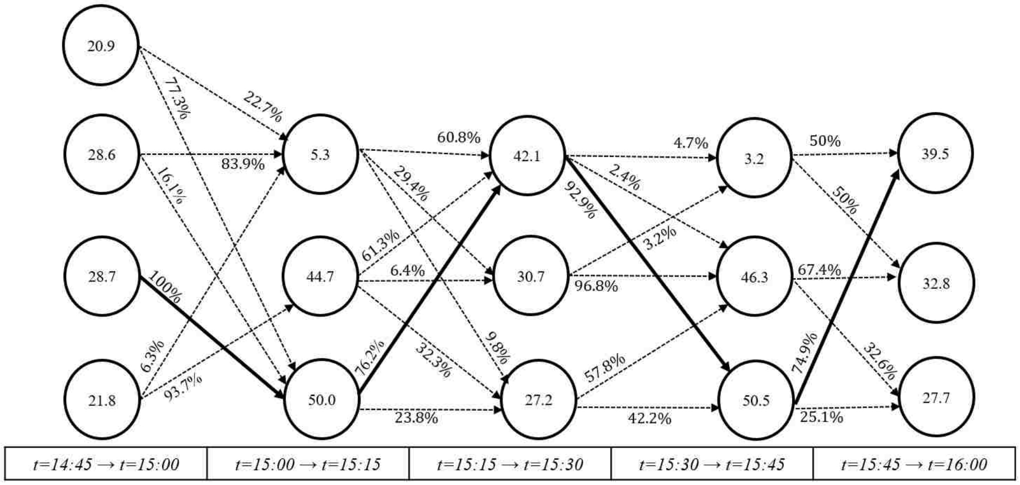


Figure 10. Transition probabilities across time intervals $t = 16{:}00$ to $t = 17{:}00$

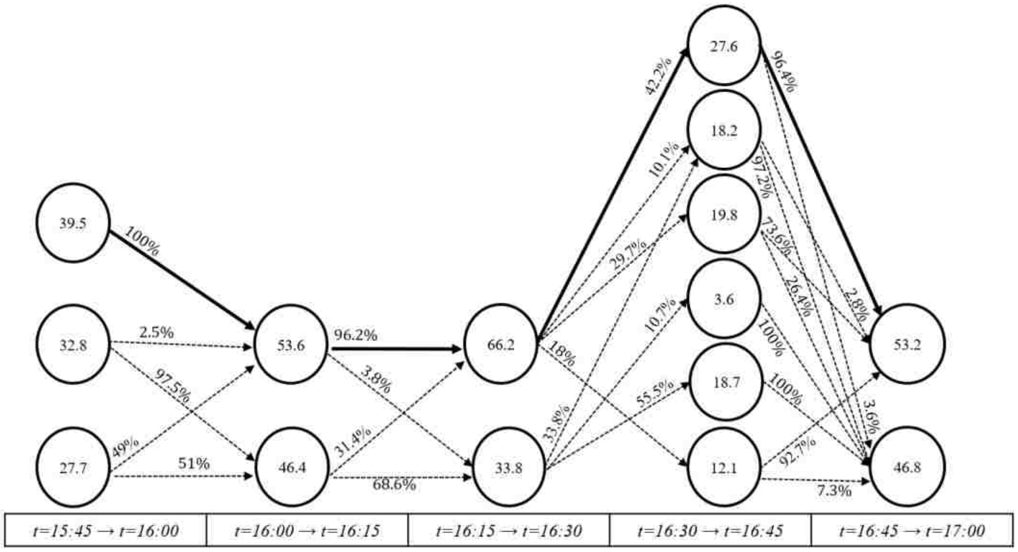


Figure 11. Transition probabilities across time intervals $t = 17{:}00$ to $t = 18{:}00$

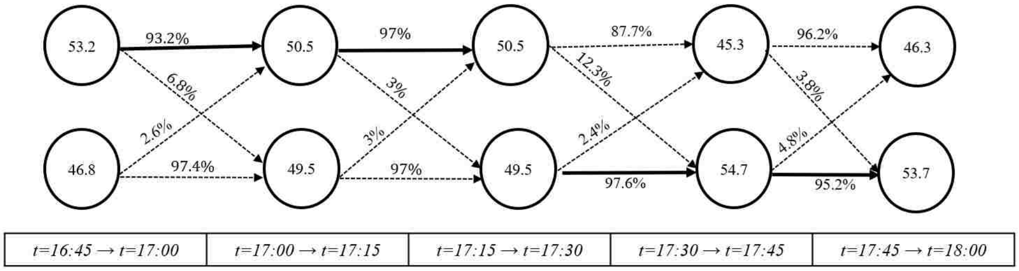
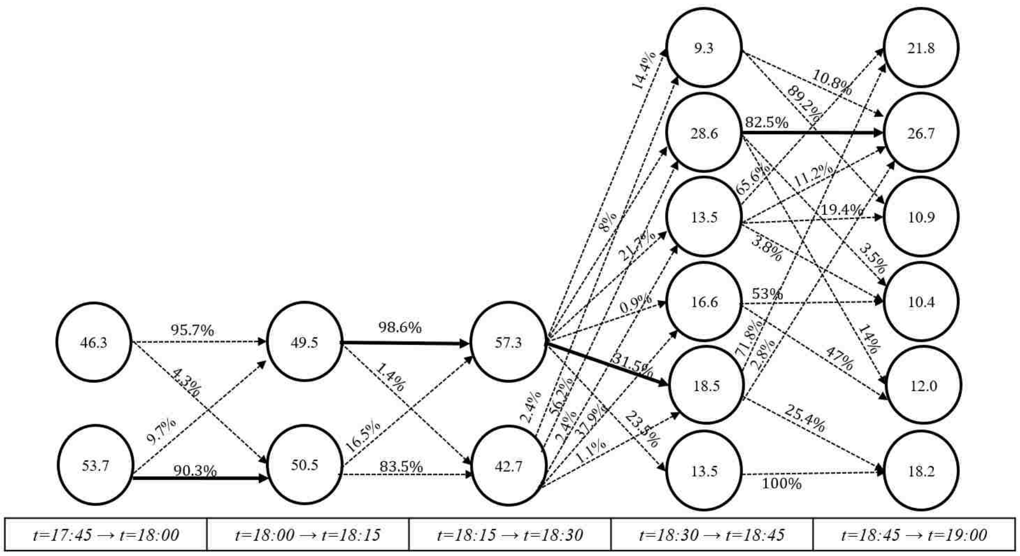


Figure 12. Transition probabilities across time intervals $t = 18{:}00$ to $t = 19{:}00$



Sections with higher variation in values of traffic characteristics

A detailed observation of the data revealed that there are a large number of sensors with low variation in the data. To observe substantial variations of traffic states, sets of sensors with large changes in traffic characteristics were grouped. For illustration purposes, two sections of the network where the values of its traffic characteristics had more variation in the traffic states were located. Figure 13 shows these sections highlighted with a red square. Section 1 is located in the intersection of the freeways, I-15 and I-515 near downtown Las Vegas. This section includes data from 52 sensors. Section 2 is located in the intersection of the freeways, I-15 and I-215. This section includes data from 35 sensors.

Traffic Characteristics at Located Section

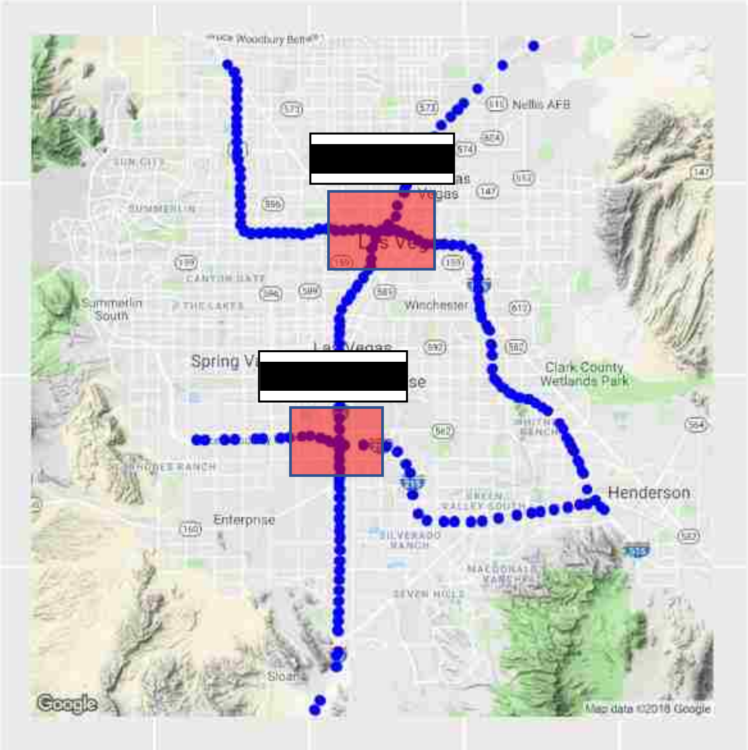
Table 5 presents the aggregated traffic characteristics of Section 1 during 96 time intervals of the 192 days considered. These results are used to obtain averages of traffic characteristics from sensors that registered similar variations and traffic conditions, the values obtained from the network-wide study averaged values of sensors with dissimilar variations and traffic conditions. The minimum speed calculated for this section is 42.71 mph at t=17:30 while the minimum speed calculated in the network wide study for the same time interval was 55.00 mph. These results evidence the existent difference between a section and a network-wide study. These results also evidence that the methodology proposed in this study can be applied to sections of the network.

Table 5. Measured traffic characteristics for Section 1

Time Interval	Avg Speed (mph)	Avg Volume (# vehicles)	Occupancy (%)	Time Interval	Avg Speed (mph)	Avg Volume (# vehicles)	Occupancy (%)
0:00	61.67	363.31	3.04	12:00	56.88	981.92	9.65
0:15	61.65	368.53	3.06	12:15	56.7	1018.37	9.93
0:30	61.64	337.51	2.8	12:30	56.35	1025.79	10.1
0:45	61.78	293.31	2.37	12:45	56.1	1031.57	10.21
1:00	61.71	251.64	2	13:00	56.17	1009.52	10.07
1:15	61.64	249.99	1.98	13:15	55.88	1049.57	10.38
1:30	61.6	235.45	1.85	13:30	54.63	1063.32	10.97
1:45	61.79	217.62	1.71	13:45	53.36	1076.52	11.64
2:00	61.58	193.07	1.48	14:00	52.95	1059.48	11.7
2:15	61.58	201.99	1.54	14:15	52.03	1099.19	12.35
2:30	61.56	199.13	1.54	14:30	49.46	1104.7	13.56
2:45	61.81	198.6	1.54	14:45	47.38	1106.29	14.6
3:00	61.75	183.39	1.38	15:00	46.95	1091.69	14.72
3:15	61.6	198.97	1.54	15:15	47	1104.16	14.78
3:30	61.96	231.94	1.83	15:30	45.92	1103.36	15.29
3:45	62.52	257.71	2.04	15:45	45.1	1100.53	15.68
4:00	62.36	238.41	1.91	16:00	45.01	1081	15.58
4:15	62.36	279.04	2.24	16:15	44.89	1103.06	15.83
4:30	62.62	370.11	3.02	16:30	44.09	1092.29	16.21
4:45	63.16	478.57	3.9	16:45	43.69	1088.11	16.37

5:00	62.93	465.04	3.79	17:00	44.22	1075.93	15.93
5:15	62.39	588.07	4.92	17:15	43.9	1098.66	16.22
5:30	60.8	797.11	7.09	17:30	42.71	1085.21	16.62
5:45	57.45	941.94	9.65	17:45	44.03	1056.21	15.67
6:00	57.4	861.61	9.06	18:00	47.3	1013.11	13.85
6:15	57.66	958.27	9.41	18:15	50.48	1019.42	12.46
6:30	54.04	1081.47	11.88	18:30	52.37	1006.59	11.45
6:45	50.41	1115.51	13.96	18:45	54.76	956.01	10.08
7:00	50.56	1071.36	13.51	19:00	57.37	881.65	8.47
7:15	51.18	1115.6	13.4	19:15	58.77	877.51	7.81
7:30	48.85	1137.65	14.8	19:30	59.27	854.52	7.42
7:45	46.52	1125.8	15.97	19:45	59.95	796.58	6.83
8:00	47.01	1075.42	15.31	20:00	60.32	742.39	6.26
8:15	49.25	1062.36	13.93	20:15	60.16	755.24	6.34
8:30	50.87	1063.27	13.07	20:30	60.04	751.14	6.29
8:45	51.68	1057.56	12.76	20:45	60.47	712.75	5.88
9:00	53.54	1017.74	11.53	21:00	60.59	675.86	5.59
9:15	55.64	1004.98	10.44	21:15	60.11	696.67	5.93
9:30	56	1020.49	10.31	21:30	59.67	684.53	6.05
9:45	55.94	1023.23	10.36	21:45	59.74	638.79	5.82
10:00	56.85	962.93	9.6	22:00	59.93	582.61	5.38
10:15	57.7	957.6	9.2	22:15	59.76	601.89	5.57
10:30	57.68	986.01	9.35	22:30	59.65	584.35	5.48
10:45	57.39	1000.14	9.59	22:45	60.12	528.92	4.94
11:00	57.49	970.69	9.37	23:00	60.44	471.37	4.36
11:15	57.52	984.13	9.44	23:15	60.6	491.57	4.44
11:30	56.91	993.77	9.69	23:30	60.61	474.11	4.3
11:45	56.71	1001.05	9.84	23:45	61.06	430.69	3.85

Figure 13. Sections with highest variation in values of traffic characteristics



Clustering and Transition Probabilities at Located Section

Figures 14 to 17 display the results obtained after applying the processes of clustering and calculation of transition probabilities to the Section 1.

Figure 14. Section 1, Transition Probabilities for time intervals $t = 15{:}00$ to $t = 16{:}00$

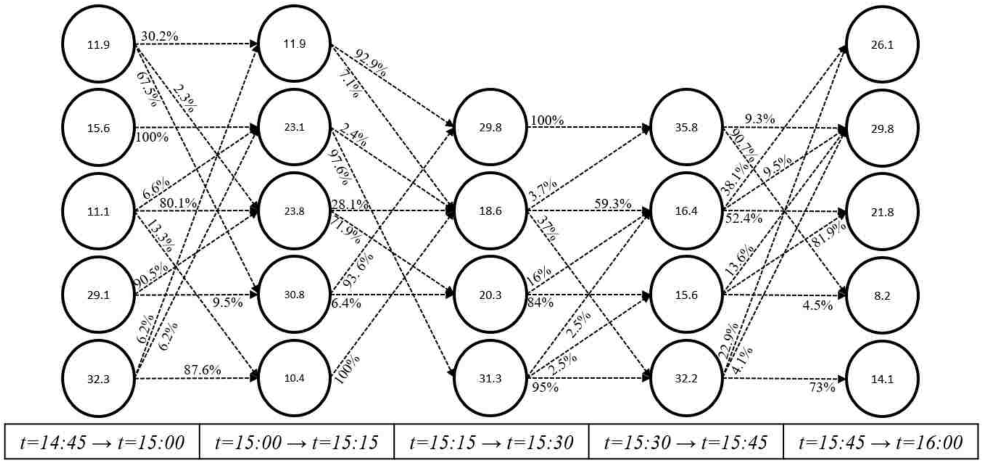


Figure 15. Section 1, Transition Probabilities for time intervals $t = 16{:}00$ to $t = 17{:}00$

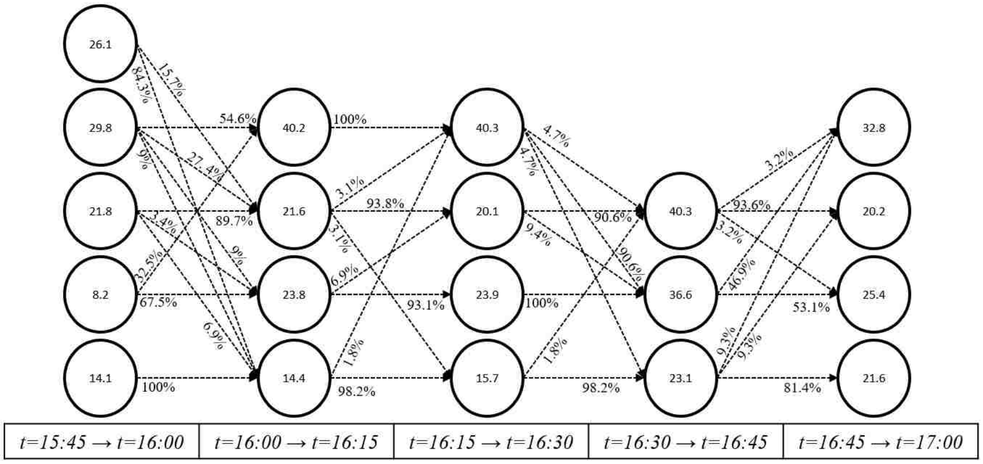


Figure 16. Section 1, Transition Probabilities for time intervals $t = 17\!:\!00$ to $t = 18\!:\!00$

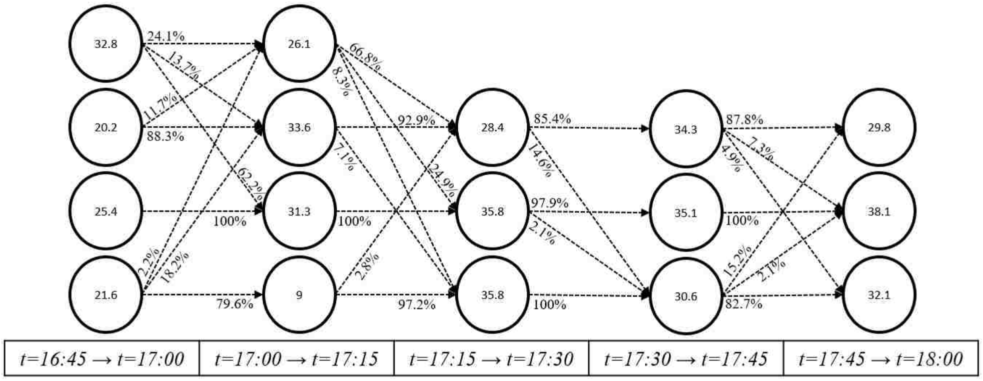
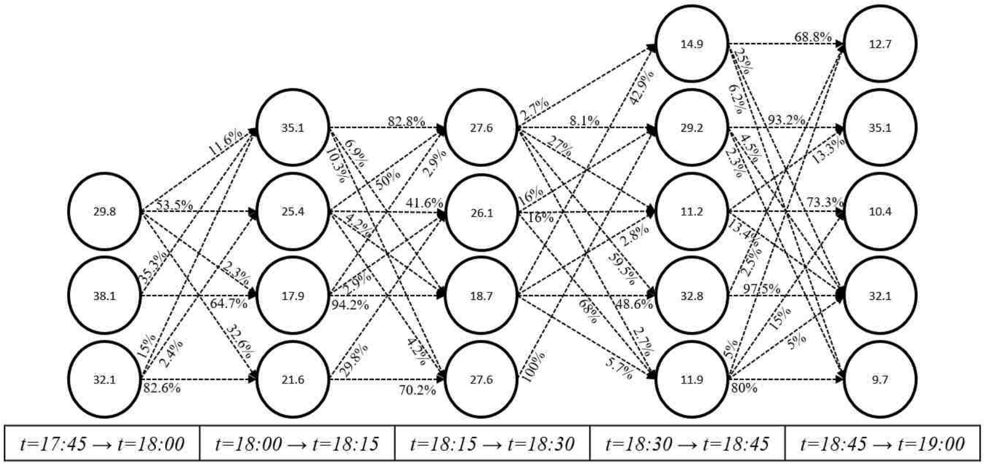


Figure 17. Section 1, Transition Probabilities for time intervals $t = 18\!:\!00$ to $t = 19\!:\!00$



Clusters and Transition Probabilities Applied to Forecast Traffic States

A system to forecast traffic states based on the obtained clusters and transition probabilities was developed. The 10-fold cross-validation method was used to test the accuracy of the forecasting system. 70% of the 192 days of the network data were used to create clusters and calculate the transition probabilities among them -training data-. The remaining 30% -testing data- was used to assess the match between the observed and the forecasted transitions between clusters. A successful forecast was considered as the match between traffic characteristics m^{yswdt} and $m^{yswdt+1}$ of the testing dataset with the clusters $K_{q(y,d,t)}^{y\mathrm{d}t}$ and $K_{q\prime(y,d,t)}^{y\mathrm{d}(t+1)}$ -respectively- that are linked by the highest transition probability from the training dataset. To apply the 10-fold cross validation method, the processes of selection of training and testing data and the forecasting process was repeated 10 times. Figure 18 illustrates the process of the forecasting system and more specifically what is considered a successful forecast. Table 6 displays the results obtained during this process, the matches are counted as the number of successful forecasts out of a total of 928 test cases. These test cases were created considering the 30% of the 192 days = 58 days, and the 16 time intervals processed in the experiments. The percentages of success are calculated as the number of successful forecasts out of the 928 test cases. A more detailed description of the average error between time intervals for all test cases can be found in Table 7.

Figure 18. Forecasting System Validation

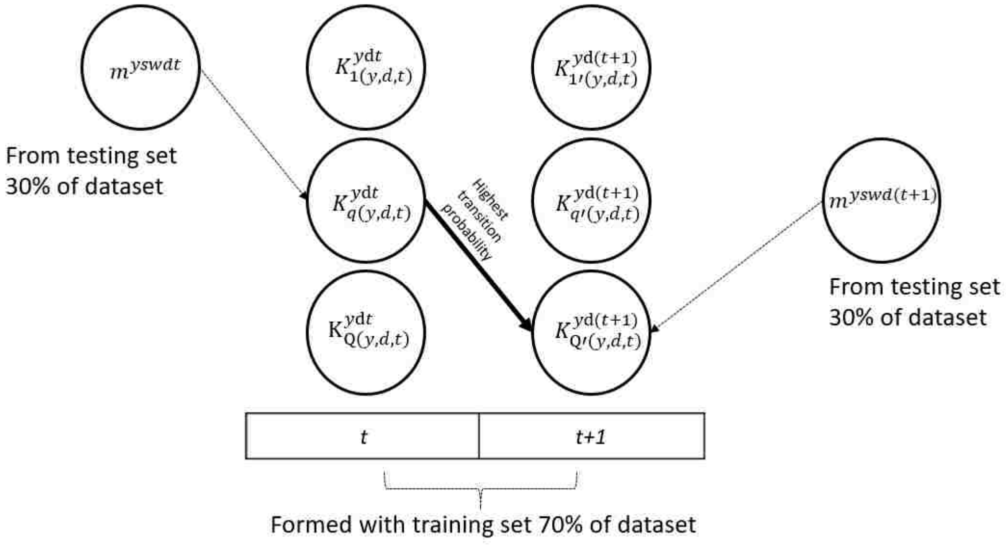

m^{yswdt}
From testing set
30% of dataset
$K_{1(y,d,t)}^{y\mathrm{d}t}$
$K_{q(y,d,t)}^{y\mathrm{d}t}$
$\mathrm{K}_{\mathrm{Q}(y,d,t)}^{ydt}$
Highest transition probability
$K_{1\prime(y,d,t)}^{y\mathrm{d}(t+1)}$
$K_{q\prime(y,d,t)}^{y\mathrm{d}(t+1)}$
$K_{\mathrm{Q}\prime(y,d,t)}^{y\mathrm{d}(t+1)}$
$m^{yswd(t+1)}$
From testing set
30% of dataset
t
t+1
Formed with training set 70% of dataset

Table 6. 10-Fold Cross-validation Results for All Time Intervals Considered

Iteration	Matches Count	Percentage of Success
1	898	96.76 %
2	880	94.82 %
3	831	89.54 %
4	879	94.71 %
5	893	96.22 %
6	881	94.93 %
7	861	92.78 %
8	877	94.50 %
9	864	93.10 %
10	840	90.51 %
	Total Avg	**93.79 %**

Table 7. Forecast Error Between Time Intervals

Time Intervals		Average Error Between Time Intervals
From	To	
14:45 -> 15:00	15:00 -> 15:15	7.86 %
15:00 -> 15:15	15:15 -> 15:30	8.74 %
15:15 -> 15:30	15:30 -> 15:45	9.64 %
15:30 -> 15:45	15:45 -> 16:00	4.99 %
15:45 -> 16:00	16:00 -> 16:15	4.46 %
16:00 -> 16:15	16:15 -> 16:30	6.68 %
16:15 -> 16:30	16:30 -> 16:45	9.84 %
16:30 -> 16:45	16:45 -> 17:00	5.47 %
16:45 -> 17:00	17:00 -> 17:15	9.45 %
17:00 -> 17:15	17:15 -> 17:30	5.47 %
17:15 -> 17:30	17:30 -> 17:45	5.77 %
17:30 -> 17:45	17:45 -> 18:00	6.78 %
17:45 -> 18:00	18:00 -> 18:15	5.79 %
18:00 -> 18:15	18:15 -> 18:30	8.89 %
18:15 -> 18:30	18:30 -> 18:45	7.88 %
18:30 -> 18:45	18:45 -> 19:00	4.56 %

Discussion

The results suggest that there are predominant traffic states with the time intervals that were considered. The dataset contains 192 network-wide traffic characteristics that were processed for each time interval. As shown in Figure 9, during the time interval starting at 15:00 and ending at 15:15, the percentage traffic characteristics is different for each cluster. For this specific time interval, 10 network-wide traffic characteristics are in the cluster displayed on the top (5.2%), 86 in the cluster displayed in the middle (44.7%), and 96 in the cluster displayed at the bottom (50%). These results show that there is a predominant traffic state and a second one with significant predominance but five percent below the most common. In addition, the cluster with the lowest predominance of 5% can be understood as a rare event in the network that cause infrequent traffic conditions; a similar trend can be observed for the rest of the time intervals. Another interesting result that can be observed in Figures 9-12 is a pattern that suggests that the majority of the predominant clusters are linked with the highest values of transition probabilities.

Figures 5 and 8 display estimated traffic states for specific time intervals. It is noticeable that heterogeneity in traffic conditions is present. There are intervals that exhibit higher variation in the number of estimated clusters as shown in Figures 9-12. This number of clusters varies from two to six. This result shows that the proposed framework significantly summarizes and reduces the effort required to analyze a big historical traffic dataset such as the one used in this study.

The results show a trend for the afternoon congested period. Before the congested period the number of clusters was higher compared to the period with more congestion. The results indicate that the number of clusters varies between two and four before the congested period and it is two for the congested period. In the transition between the non-congested and congested periods there was found a large number of clusters. This high variation can be attributed to commuters looking for alternate routes before the most congested periods. As expected, once the congested period is reached, less variability occurs probably due to capacity constraints. These capacity constraints minimize the choices that drivers have, reducing the effect of randomness due to human behavior. When the congested period ends, larger numbers of clusters are observed. This can be attributed to both drivers looking for alternate routes and low congestion observed during night time periods. These observed patterns can help to determine the duration of the periods with more congestion in the network. This provides a better understanding of the congested period at a regional level. This type of results can be used to develop more accurate regional travel demand models. Currently a fixed congested period is suggested, for example, in the case of the U.S., the FHWA states that the congested period can be assumed to be from 4p.m to 7p.m during weekdays (FHWA, Urban Congestion Report., 2018). However, as shown in this study, the afternoon congested period for Las Vegas area was found to start at 5:00p.m. and end at 6:30p.m.

In the results obtained from the studied section -compared to the results at network level- a higher variation of the values of the traffic characteristics can be observed. The number of sensors considered for the section study (52) are significantly less than the ones considered for the study of the entire network (466). This enabled obtaining averaged values from a smaller number of traffic characteristics; from sensors that present more common traffic characteristics during each time interval. These results can contribute to allocate and obtain values of the traffic characteristics of congested sections in the network.

CHAPTER 5 CONCLUSIONS, FUTURE WORK AND LIMITATIONS

This study proposes a framework to capture dynamic traffic trends at a network-wide level using sensor data. A mathematical programming formulation and solution algorithm are proposed to generate average traffic states and transition probabilities among them. A k-means algorithm was used to generate clusters of data which centroid represents a traffic state. Multiple centroids could be observed for a time interval; each representing homogeneous traffic flow dynamics. The availability of these centroids minimizes the data processing effort required for traffic analyses at the network level; usually, traffic datasets include hundreds of thousands of records. Transition probabilities were computed considering the frequency of an event. A historical dataset for Las Vegas freeways was used in this study.

The proposed framework is able to reveal predominant traffic conditions for each time interval. The frequency of a traffic condition is given by the number of network-wide traffic characteristics within a cluster; in other words, how many of the total number of analyzed days experienced a specific traffic condition. Knowing the frequency and predominance of traffic conditions can contribute to anticipate and prepare the network for such conditions. This result is valuable for transportation planning and congestion studies. An important insight from the analysis is a better understanding and data driven observation of a congested period at a regional level.

The proposed framework enabled the creation of a forecasting system, this system was developed to illustrate one of the possible usages of the obtained clusters and transition probabilities. In this system a successful forecast was considered as a match between the observed and the forecasted transitions between the clusters at subsequent time intervals. The process was validated with the 10-fold cross validation method helping to ensure the consistency of the proposed forecasting system. After validating the results, the forecasting system reached an average accuracy of 93.79%.

The proposed analysis framework has multiple potential applications that require further research. For example, the ability to estimate future traffic states based on observed real-time data enables the generation of anticipatory traffic management strategies such as route guidance and demand responsive traffic control. The clusters and transition probabilities generated by the proposed framework can be used to perform short-term forecasting. Real-time network conditions can be used to determine its cluster membership. Then, the traffic state of the following time intervals can be estimated using the highest corresponding transition probabilities associated with the cluster under default traffic control or information conditions. The current work does not factor for working zones and the validation of the existence of clusters that represent traffic states with incidents make part of future work.

REFERENCES

1. Padiath, A., L. Vanajakshi, and S. Subramanian. Estimating Spatial Traffic States with Location-Based Data Under Heterogeneous Conditions. *Transportation Research Record: Journal of the Transportation Research Board*, 2012. 2291: 72-79.
2. Ma, T., Z. Zhou, and B. Abdulhai. Nonlinear multivariate time-space threshold vector error correction model for short term traffic state prediction. *Transportation Research Part B: Methodological*, 2015. 76: 27–47.
3. Hashemi, H., and K. Abdelghany. Real-Time Traffic Network State Prediction for Proactive Traffic Management. *Transportation Research Record: Journal of the Transportation Research Board*, 2015. 2491: 22–31.
4. Paz, A., and S. Peeta. Behavior-consistent real-time traffic routing under information provision. *Transportation Research Part C: Emerging Technologies*, 2009. 17: http://dx.doi.org/10.1016/j.trc.2009.05.006.
5. Dimitrakopoulos, G., and P. Demestichas. Intelligent Transportation Systems. *IEEE Vehicular Technology Magazine*, 2010. 5.1: 77–84.
6. Zavin, A., A. Sharif, A. Ibnat, W. M. Abdullah, and M. N. Islam. Towards developing an intelligent system to suggest optimal path based on historic and real-time traffic data. *Computer and Information Technology, 2017 20th International Conference of IEEE*. http://ieeexplore.ieee.org/document/8281834/. Accessed Jul. 20, 2018.
7. Dogru, N., and A. Subasi. Comparison of clustering techniques for traffic accident detection. *Turkish Journal of Electrical Engineering & Computer Sciences*, 2015. 23: 2124–2137.
8. Paz, A., and S. Peeta. Information-based network control strategies consistent with estimated driver behavior. *Transportation Research Part B: Methodological*, 2009. 43: 73–96.
9. Kachroo, P., N. Shlayan, A. Paz, S. Sastry, and S. K. Patel. Model-Based Methodology for Validation of Traffic Flow Detectors by Minimizing Human Bias in Video Data Processing. *IEEE Trans. Intelligent Transportation System*, 2015. 16: 1851–1860.
10. Vlahogianni, E. I., M. G. Karlaftis, and J. C. Golias. Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, 2014. 43: 3–19.
11. Nemtanu, F. C., I. M. Costea, I. Badescu, V. Iordache, and J. Schlingensiepen. The framework of using models for comparative assessment of traffic sensors. *Design and Technology in Electronic Packing, 2016 IEEE 22nd International Symposium for IEEE*, 2016. 200–204.
12. Van Lint, J. W. C., and S. P. Hoogendoorn. A Robust and Efficient Method for Fusing Heterogeneous Data from Traffic Sensors on Freeways. *Computer-Aided Civil and infrastructure Engineering*, 2010. 25: 596–612.
13. Koroliuk, M., and C. Connaughton. Analysis of big data set of urban traffic data. *M2 Project (30 ECTS)2015, Project Report*, 2015. 1–12.
14. Paz, A., and Y.-C. Chiu. Adaptive Traffic Control for Large-Scale Dynamic Traffic Assignment Applications. *Transportation Research Record: Journal Transportation Research Board*, 2011. 2263: 103–112.
15. Gu, H. A Probabilistic Approach for Traffic State Estimation. *2016 IEEE 19th International Conference on Intelligent Transportation Systems*, 2016. 1: 2595–2600.

16. Zhang, Y., N. Ye, R. Wang, and R. Malekian. A Method for Traffic Congestion Clustering Judgment Based on Grey Relational Analysis. *ISPRS International Journal of Geo-Information*, 2016. 5: 71.

17. Bharadwaj, H. S., S. Biswas, and K. R. Ramakrishnan. A Large Scale Dataset for Classification of Vehicles in Urban Traffic Scenes. *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing*, 2016. 16: 1–8.

18. Zhao, J. Research on Prediction of Traffic Congestion State. *MATEC Web of Conferences*, 2015. 9: 1-5.

19. Xu, J., D. Deng, U. Demiryurek, C. Shahabi, and M. Van Der Schaar. Mining the Situation: Spatiotemporal Traffic Prediction with Big Data. *IEEE Journal Selected Topics in Signal Process*, 2015. 9: 702-715.

20. Allström, A, J. Ekstrom, D. Gundlegard, R. Ringdahl, C. Rydergren, A.M. Bayen, and A.D. Patire. Hybrid Approach for Short-Term Traffic State and Travel Time Prediction on Highways. *Transportation Research Record: Journal of the Transportation Research Board*, 2016. 2554: 60-68.

21. Oh, S., Y. J. Byon, and H. Yeo. Improvement of Search Strategy with K-Nearest Neighbors Approach for Traffic State Prediction. *IEEE Transaction on Intelligent Transportation System*, 2016. 17.4: 1146–1156.

22. Barros, J., M. Araujo, and R. J. F. Rossetti. Short-term real-time traffic prediction methods: A survey. *Models and Technologies for Intelligent Transportation System,* 2015 International Conference, 2015. June: 132–139.

23. Paz, A., and S. Peeta. Paradigms to deploy a behavior-consistent approach for information-based real -time Traffic routing. *Networks and Spatial Economics*, 2009. 9.2: 217–241.

24. Paz, A., and S. Peeta. On-line calibration of behavior parameters for behavior-consistent route guidance. *Transportation Research Part B Methodological*, 2009. 43. 4: 403–421.

25. Paz, A., and S. Peeta. Fuzzy control model optimization for behavior-consistent traffic routing under information provision. *IEEE Transactions on Intelligent Transportation System*, 2008. 9.1: 27–37.

26. Lu, F., Y. Duan, and N. Zheng. A practical route guidance approach based on historical and real-time traffic effects. *2009 17th International Conference in Geoinformatics*, 2009. 2006.

27. Boriboonsomsin, K., M. J. Barth, W. Zhu, and A. Vu. Eco-routing navigation system based on multisource historical and real-time traffic information. *IEEE Transactions on Intelligent Transportation System*, 2012. 13. 4: 1694–1704.

28. Xu, Y. Stochastic Traffic Control based on Regional State Transition Probability Model. *Service Operations and Logistic, and Informatics, 2016 IEEE International Conference*. Jul. 10: 89–94.

29. Dou, H., G. Wang, and M. Guo. Traffic Guidance Oriented Model of Traffic State Probability Forecast. *Journal of Transportation System Engineering and Information Technology*, 2011. 11.2: 27–32.

30. Noroozy, R., and B. Hellinga. Real-time Prediction of Near-Future Traffic States on Freeways Using a Markov Model. *Transportation Research Record: Journal of the Transportation Research Board*, 2014. 2421: 115–124.

31. Maheshwari, P., P. Kachroo, A. Paz, and R. Khaddar. Development of control models for the planning of sustainable transportation systems. *Transportation Research Part C: Emerging Technologies*, 2015. 55: 474–485.

32. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 1987. 20. C: 53–65.

33. Nawrin, S. Exploreing K-Means with Internal Validity Indexes for Data Clustering in Traffic Management System. *International Journal of Advanced Computer Sciences and applications.* 2017. 8. 3.

34. Montazeri-Gh, M., and A. Fotouhi. Traffic condition recognition using the k-means clustering method. *Scientia Iranica*., 2011.18.4 B: 930–937.

35. Yingqiu, L., L. Wei, and L. Yunchun. Network Traffic Classification Using K-means Clustering. *Computer and Computational Sciences, IMSCCS 2007. Second International Multi-Symposium on IEEE*, 2007. Aug. 13: 360–365.

36. Freeway & Arterial System of Transportation. Traffic Dataset, 2016-2017. Accessed Aug. 1, 2018.

37. Federal Highway Administration. *Urban Congestion Report*. https://ops.fhwa.dot.gov/perf_measurement/ucr/documentation.htm. Accessed Aug. 1, 2018.

CURRICULUM VITAE

Graduate College
University of Nevada, Las Vegas

Carlos Gaviria
carlos.gaviria.engineer@gmail.com

Degrees:

Bachelor in Computer Science, 2015
Universidad del Cauca, Colombia

Publications:

- Paz A., Gaviria C., Arteaga C. and Jose Torres-Jimenez. (2018). Mining Dynamic Network-Wide Traffic States. In 21st IEEE International Conference on Intelligent Transportation Systems.
- Paz A., Arteaga C. & Gaviria C. (2018). Integrated System for Collecting and Reporting Crash and Citation Data. In Proceedings of the 4th International Conference on Vehicle Technology and Intelligent Transport Systems - Volume 1: VEHITS, ISBN 978-989-758-293-6, pages 225-230. DOI: 10.5220/0006648202250230.
- Cobos, C., Erazo, C., Luna, J., Mendoza, M., Gaviria, C., Arteaga, C., & Paz, A. (2016, September). Multi-objective Memetic Algorithm Based on NSGA-II and Simulated Annealing for Calibrating CORSIM Micro-Simulation Models of Vehicular Traffic Flow. In Conference of the Spanish Association for Artificial Intelligence (pp. 468-476). Springer International Publishing.
- Cobos, C., Daza, C., Martínez, C., Mendoza, M., Gaviria, C., Arteaga, C., & Paz, A. (2016, November). Calibration of Microscopic Traffic Flow Simulation Models Using a Memetic Algorithm with Solis and Wets Local Search Chaining (MA-SW-Chains). In Ibero-American Conference on Artificial Intelligence (pp. 365-375). Springer International Publishing.
- Paz, A., Martinez, E., Molano, V., Gaviria, C., & Arteaga, C. (2015). Calibration of Traffic Flow Models Using a Memetic Algorithm. Transportation Research Part-C: Emerging Technologies, 55, 432-443. doi: 10.1016/j.trc.2015.03.001.
- Paz, A., Molano, V., & Gaviria, C. (2012). Calibration of CORSIM Models Considering all Model Parameters Simultaneously. 15th International IEEE Conference on Intelligent Transportation System (ITSC), Anchorage, AK.

Thesis Title:

A FRAMEWORK TO CAPTURE DYNAMIC TRAFFIC TRENDS FROM
HISTORICAL SENSOR DATA

Thesis Examination Committee:

Chairperson, Alexander Paz, Ph.D.
Committee Member, Dave James, Ph.D.
Committee Member, Mohamed Kaseko, Ph.D.
Graduate College Representative, Justin Zhan, Ph.D.