Summer 2015

Shape based classification and functional forecast of traffic flow profiles

Wasim Kayani

Department: Civil, Architectural and Environmental Engineering

SHAPE BASED CLASSIFICATION AND FUNCTIONAL FORECAST OF TRAFFIC FLOW PROFILES

by

WASIM KAYANI

A DISSERTATION

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

in

CIVIL ARCHITECTURAL AND ENVIRONMENTAL ENGINEERING

2015

Dr. Ivan Guardiola, Advisor
Dr. V A Samaranayake
Dr. Cesar Mendoza
Dr. Mohamed ElGawady
Dr. J David Rogers

ABSTRACT

This dissertation proposes a methodology for traffic flow pattern analysis, its validation, and forecasting. The shape of the daily traffic flows are directly related to the commuter's traffic behavior which merit analysis based on their shape characteristics. As a departure from the traditional approaches, this research proposed a methodology based on shape for traffic flow analysis. Specifically, Granulometric Size Distributions (GSDs) were used to achieve classification of daily traffic flow patterns. A mathematical morphology method was used that allows the clustering of shapes. The proposed methodology leads to discovery of interesting daily traffic phenomena such as five normal daily traffic shapes beside abnormal shapes representing accidents, congestion behavior, peak time fluctuations, and malfunctioning sensors.

To ascertain the significance of shape in traffic analysis, the proposed methodology was validated through a comparative classification analysis of the original data and GSD transformed data using the Back Prorogation Neural Network (BPNN). Results demonstrated that through shape based clustering more appropriate grouping can be accomplished that can result in better estimates of model parameters.

Lastly, a functional time series approach was proposed to forecast traffic flow for short and medium-term horizons. It is based on functional principal components decomposition to forecast three different traffic scenarios. Real-time forecast scenarios of partially observed traffic profiles through Penalized Least squares (PLS) technique were also demonstrated. Functional methods outperform the conventional ARIMA model in both short and medium-term forecast horizons. In addition, performance of functional methods in forecasting beyond one hour was also found to be robust and consistent.

ACKNOWLEDGMENTS

I would like to give sincere thanks to my adviser, Dr. Ivan Guardiola. His insights, guidance, expertise and patience helped me through every step of my PhD. For sharing his extensive knowledge, and for always challenging and bringing best out of me. I would also like to thank Dr. Samaranayake for introducing me to the field of statistics and helping me build my research foundation. I am thankful to Dr. Cesar Mendoza, Dr. Muhamad ElGawady and Dr. David Rogers for agreeing to be part of my PhD defence committee as well as for their valuable input during my comprehensive examination.

My parents deserve much credit for their love and encouragement through every stage of my life. They have provided invaluable support and help at every turn to pursuit my academic goals. Also, my brothers and sisters have all set a good example for me to follow. My Kids Azm, Kamila and Wajiha who supported quietly during my studies.

Above all, my wife Zahra who helped me through this incredible journey and inspired me to never stop striving for my potential. I feel privileged to have you by my side everyday. I also want to thank everyone in the lab especially Sidharat, Brian and Shikhar who were always helpful to me in solving research riddles.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Table Page

1. INTRODUCTION, MOTIVATIONS, AND LITERATURE REVIEW

Traffic flow modeling is an essential component of any traffic control, monitoring, and management system. It forms the basis for ITS technologies, like Advanced Traveler Information Systems (ATIS) and Advanced Traffic Management Systems (ATMS) as examples, which attempt to deal with the traffic congestion, accidents and travel time reliability problems. The overall objective of these systems is to increase the operational efficiency and capacity of the transportation network. This objective of achieving efficient transportation systems is only possible through sound traffic analysis based on traffic flow theory. Traffic flow is influenced by travel's behavior as well as abrupt disturbances because of various unexpected events (e.g., accidents, weather-induced disruption) that may change the underlying dynamics and the stability of the data generation process. It implies that daily traffic flow profiles exhibits unique shapes that can be grouped into definite patterns. To explore these patterns, clustering is considered a well established technique already employed in traffic domain. The patterns obtained through clustering can assist in understanding traffic dynamics and then help in simulation and forecasting etc. The shape of the traffic profiles also suggest that traffic data is functional in nature and therefore suitable for application of function data analysis techniques for forecasting.

Traffic flow theory provides the basis for traffic analysis. It helps to understand the traffic dynamics as well as to develop mathematical relationships among the primary elements of the traffic stream: flow, density, and speed. Among these traffic parameters, as suggested by the name “Traffic Flow Theory”, flow is the primary element which best describes the traveller's behavior. It is therefore the focus of modern Intelligent Transportation Systems research and practice. Traffic flow can be considered as both a temporal and a spatial phenomenon. Its conceptual as well as practical dependence on time and space has been well established in the traffic forecasting literature.

Traffic flow analysis is done at three different levels or through models namely : microscopic, macroscopic, and mesoscopic (Boxil et al, 2005). Microscopic models predict the individual following behavior of cars (change in speed and position) as a function of the behavior of the leading vehicle. Wiedemann model, Intelligent driver model and Gipps' model are all examples of car following models (microscopic models). Macroscopic traffic flow theory relates to traffic flow, running speed, and density. Analogizing traffic to a stream has principally been developed for limited access roadways (i.e expressways) (Leutzbach, 1988). The most widely used model is the Greenshields model, which demonstrates that the relationships between speed and density is linear (Erlingsson et al, 2006). Finally, Mesoscopic (kinetic) models are those that combine characteristics of both the microscopic and macroscopic levels. Macroscopic properties like flow and density are the product of individual (microscopic) decisions. Yet those microscopic decision-makers are affected by the environment around them, (e.g. the macroscopic properties of traffic). Hence, macroscopic level is appropriate for traffic analysis where the focus is on the traffic stream rather than on individual vehicles.

Daily traffic flow is a function of time and defines the functional behavior of commuters over time. The basic traffic stream modeling relationship relates flow (q) to the product of density (k) and space mean speed (v) as $q = kv$. However, this daily traffic flow not only varies within a day but also differs within the weekdays. Figure 1.1 illustrates three time series curves of traffic flow. This figure illustrates that there is significant changes in daily traffic flow profile shapes throughout the week, which are influenced by a variety of factors. Refer Figure 1.1, the Wednesday traffic profile shape represents a typical bi-modal (typical morning and afternoon peak) working day behavior. Sunday represents a uni-modal shape which is typical non-working day behavior. While the Saturday profile illustrated is a non-working day, depicts a very unique traffic pattern, which was highly influenced by a major weather event (e.g. a blizzard). The figure also indicates the presence of some patterning in the flow by time of day and day-to-day. However, simple visualization of the data is often not satisfactory in determining if there is similarity or dissimilarity from one day to another. Hence, a more analytical approach to determine the qualitative characteristics of the data is to undertake clustering analysis as it is helpful in identifying, or recognizing patterns of interest within a historical data set.

Clustering enables to derive traffic patterns systematically using matching criteria contrary to the practice of predefined days into classes. It result in obtaining site specific adaptive sets of traffic patterns. The traffic Monitoring Guide (TMG) (FHWA2001) identifies this necessity and also acknowledges clustering analysis as appropriate technique for this purpose.

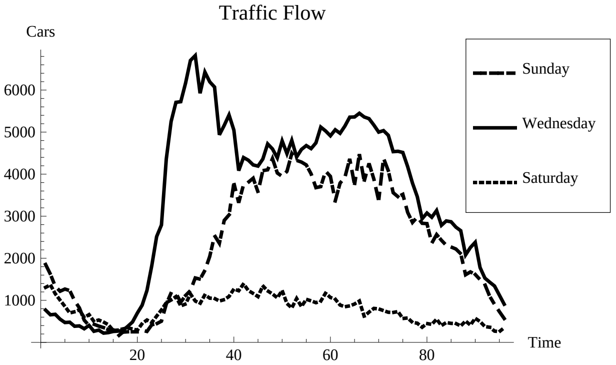


Figure 1.1. Daily traffic flow time series for Sunday-11/7/10, Wednesday- 11/10/10 and Saturday-12/11/10.

Traffic flow patterns have numerous uses. The most important use could be as a planning tool, where they lay the basis for setting management scenarios (e.g in planning, one might wish to estimate the annual traffic volume over the planned horizon for proposed infrastructure alternatives). The annual volume is then used for estimating the expected saving in travel time for economic feasibility studies. Similarly for design purposes, hourly traffic volume is often required to determine the facilities capacity. Thus, accurately predicting the hourly flow variations would become essential to avoid an over-design or under-design of new facilities. By exploring spatiotemporal traffic patterns, more insightful information may provide us an understanding of freeway traffic that can be used for effective traffic management, traffic control, organization, and other engineering applications, which should increase freeway capacity, improve traffic safety, and result in high-quality mobility. In particular, surveying the congested traffic patterns could give us necessary information for efficient collective management strategies, including such well-know methods as ramp metering and traffic assignment (Lan et al, 2008). (Varaiya, 2005) argues that effective management on highway congestion through investigation of traffic patterns can significantly reduce congestion.

Traffic patterns can be also be used as input for macroscopic traffic simulators to assess the expected performance of future infrastructure modifications and take pro-active measures in advance. Another use of the obtained patterns is reconstruction of traffic data in the case of sensors malfunctioning or as a valuable reference in incident detection algorithms (Weil et al, 1998). From prediction perspective, the clustering results can be used to obtain medium and long-term traffic patterns (i.e one day to one year in advance), as these demand patterns are a synthesis of the recurrent behavior of travel demand on a specific infrastructure based on historical data (Soriguera, 2012).

In the traffic classification domain, previous studies used non-shape based classification to analyze traffic patterns such as the works/research of (Rakha and Van, 1995), (Wild, 1997) and (Chung, 2003). In (Rakha and Van, 1995), two groups were determined which are {Tuesdays, Wednesdays, Thursdays} and {Mondays, Fridays, Saturdays, Sundays}. Furthermore, (Wild, 1997) classified week days into six groups. The results of these studies vary and thus literature lacks consistency regarding classification of days into a clear set or number of groups. (Weijermars et al, 2007) was the first to suggest the need for study the shape of the traffic flow profiles. The study concluded that shape of the daily flow profiles may differ between different type of days and motivates the need to such an analysis of traffic patterns based on their inherent shape characteristics.

In Chung *et al.* (2001), agglomerative hierarchal clustering was performed on traffic flow to develop a short-term prediction model, where clustering was used to develop a more precise model. However, he concludes, that his model which uses a historical average can capture the shape of the historical traffic pattern, but the model is highly sensitive to the presence of outliers within the data, which are a result of significant weather events and other external factors commonly found in real-world data. This emphasizes the importance of shape as anomalies in the historical data can cause the models to deviate or become ineffective which is quite common (Rakha and Van (1995)). Examples of such models are the numerous time series based models, which under perform in predicting the time horizon, refer to (Hogberg, 1976; Ahmed and Cook, 1979; Ahmed, 1983; Okutani and Stephanedes, 1984; Stephanedes et al, 1990).

Traffic flow forecasting is an essential part of transportation planning, traffic control, and intelligent transportation systems (Tan et al, 2009). Most of the research effort is devoted to the short-term traffic forecasting (5 - 30 minutes) due to its usefulness in real time applications. However, medium and long-term (several hours to days and weeks) forecasting is also significant in terms of its utility towards planning purposes. (Jiang et al, 2005) highlighted that for long-term forecasting, it should be understood that traffic flow is highly complex and not amenable to accurate mathematical modeling. Therefore, nonparametric methods and adaptive algorithms are required to learn and recognize patterns in an effective manner.

The literature on traffic forecasting reveals that mainly three approaches are used for predictive traffic analysis: neural networks (NNs); neighbor nonparametric regression and autoregressive integrated moving average (ARIMA) time series models. The first two approaches are non parametric in nature while ARIMA is a parametric technique. However, keeping the functional nature of data, a non parametric technique seems more appropriate for traffic analysis. The literature further emphasizes that out of these tools neural networks is a convenient and effective tool for developing relationships between streams of input and output data, not only for pattern recognition to which they are usually associated, but also for a wide range of modeling situations (Kirby et al, 1997). Kirby et al 1997 used a BPNN to carry out a comparative study on NN and statistical models such as regression. Neural Networks have performed better than contemporary statistical techniques like discriminant analysis, negative binomial regression, stepwise logistic regression and other classical techniques used in incident detection methodological development (Ivan and Sethi, 1998), (Khan and Ritchie, 1998), gap acceptance modeling (Pant and Balakrishnan, 1994), and safety modeling (Hashemi et al, 1995) , (Chang, 2005), (Sommer et al., 2008).

In contrast to classical statistics where the focus is on a sample of data points or vectors, functional data analysis focuses on a sample of functional observations (Ramsey and Silverman, 2005), like curves and images. Functional data usually reflects the influence of certain smooth functions that are assumed to underlie and to generate the observations. Classical multivariate statistical methods may be applied to such data, but they cannot take advantage of the additional information implied by the smoothness of the underlying functions. Functional data analysis can often extract additional information contained in the functions and their derivatives, not usually available through traditional methods (Daniel et al., 2007). In-spite of its usefulness, however, this emerging field mostly remains unexplored in the traffic domain.

In traffic studies, daily traffic profiles or curves are treated as multivariate data as they are given as a finite discrete time series. This traditional multivariate approach completely ignores vital information about the smooth functional behavior of the generating process that defines the data (Green and Silverman, 1994). The basic idea behind functional shape analysis is to express discrete time series observations in the form of a function that represents the entire measured function as a single observation or datum. Later, modeling and/or prediction information is drawn from a collection of functional data by applying statistical concepts from multivariate data analysis.

The technological developments in traffic data collection evolved over recent decades, that has allowed more dense sampling of observations over time and space. Although presently classical multivariate statistical techniques are applied to such traffic data, they do not take advantage of additional information that could be gained by the smoothness of underlying functions. In particular, functional data analysis can often extract additional information contained in the function and its derivatives that is not normally available from the application of traditional statistical methods. Because the approach essentially treats the whole traffic profile (curve) as a single datum, there are correlations only between curves rather than between repeated measurements. This represents a change in philosophy towards the handling of traffic time series data and provides a motivation to consider shape and functional approaches for traffic analysis.

This research seeks to fill the gaps prevailing in the literature by introducing a shape based clustering technique, which does not require any pre-classification of existing patterns in the data. The advantage of the proposed methodology eliminates bias of heuristic preprocessing such as those found in (Wild, 1997) . In addition, the proposed methodology reduces data dimensions, reduces the computational cost (due to smaller number of observations required to describe a daily profile) and is also robust to fluctuations in flows where shape is preserved. It is effective in capturing outliers into separate groups, which provides a strong basis for precise prediction modeling and relevant research effort within traffic contexts. Above all, traffic profiles are clustered based solely on their shapes and therefore making trivial the requirement of analyzing different flow characteristics: total flow, peak flow, peak period and off peak time separately.

Understanding the benefits of shape analysis in the traffic domain, the literature search for an effective shape analysis methodology, led to Mathematical Morphology, which is regarded as an efficient technique for shape analysis in image processing. The technique was originally developed by Matheron and Serra (Serra, 1982) at the Ecole des Mines in Paris. Mathematical morphology mostly deals with the mathematical theory of describing shapes using set theory. It provides a number of useful tools for image analysis, which is based on the assumption that an image consists of structures that may be handled by set theory. Conventional image processing methods for analysis of shapes of objects requires a binary image and a subsequent calculation of factors roundness, shape, and area (Gonzlez and Woods, 1996). The central idea of these techniques is to examine the geometric structures in an image overlaying them with small patterns called structuring elements. Of all mathematical morphology analysis techniques, the most appropriate tool for discriminating shapes is the Granulometric Size Distribution (GSD) (Ballarin and Valentinuzzi, 2001) and (Vincent and Dougherty., 1994).

The GSD method can be naturally applied to the analysis of signals. In Gaston-Romeo et al (2011), GSD is applied to signals in an effort to analyze daily solar radiation time series curves. Similarly, in Guardiola and Mallor (2013), unintended electromagnetic emissions from wireless communication devices are also analyzed using GSD. In both previously mentioned works curves are identified by their subgraph and are considered to be bi-dimensional images on which morphological operators are applied to gain information regarding their shape.

In the background of above, the research is organized as follows. The section 2 *"ON TRAFFIC FLOW PATTERN SHAPE CLASSIFICATION AND ANALYSIS"*, demonstrates use of Mathematical Morphology (MM) tools to identify functional shapes, which is achieved through classification of daily traffic profiles. Specifically, the use of the Granulometric Size Distribution (GSD) is emphasized for exploring traffic patterns. Through the employment of MM, the development of an analysis of shape is carried out in an effort to generate interpretable classification of historical daily traffic patterns. The development of unique granulometries for each day are contracted and are used to cluster days into distinct groups. The section 3 *"A HYBRID OF COMPUTATIONAL INTELLIGENCE TECHNIQUES FOR SHAPE ANALYSIS OF TRAFFIC FLOW CURVES"* highlights and validates the use of shape analysis using Mathematical Morphology tools as a means to develop meaningful clustering of historical data. A comparative classification analysis of original data and GSD transformed data is carried out. The section 4 "A FUNCTIONAL TIME SERIES APPROACH FOR TRAFFIC FORECASTING" adopts a functional time series approach to forecast traffic flow for short and medium-term horizons. A technique based on functional principal components is used to forecast different traffic scenarios. Although the technique is capable of forecasting a complete day, the research focus remain on

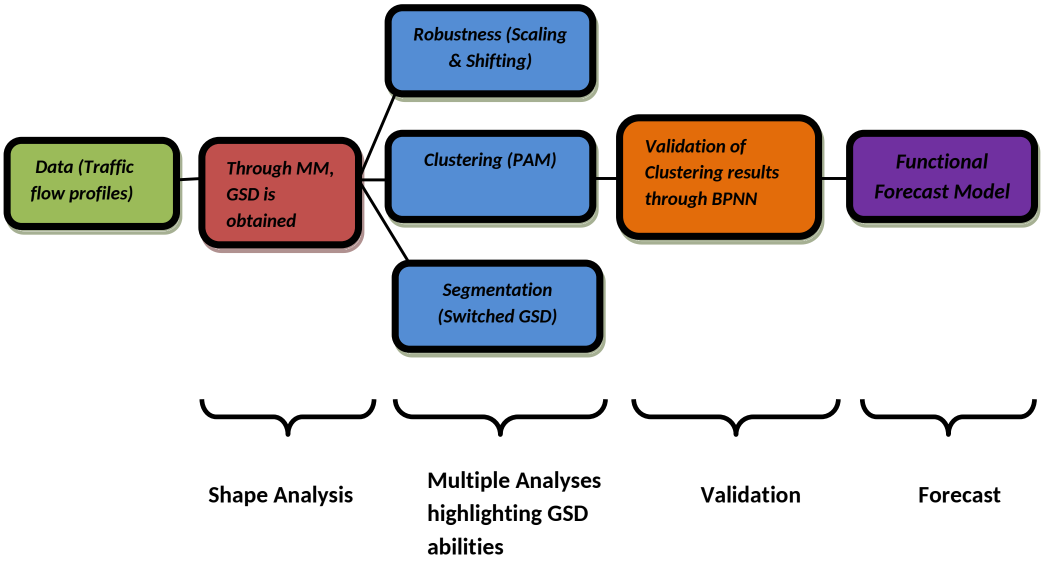

Robustness (Scaling & Shifting)
Data (Traffic flow profiles)
Through MM, GSD is obtained
Clustering (PAM)
Validation of Clustering results through BPNN
Functional Forecast Model
Segmentation (Switched GSD)
Shape Analysis
Multiple Analyses highlighting GSD abilities
Validation
Forecast

Figure 1.2. Research overview.

forecast one hour ahead. In addition, forecast scenarios for partially observed traffic profiles through Penalized Least squares (PLS) technique is also demonstrated. Results obtained are compared with a traditional benchmark, Auto Regressive Integrated Moving Average (ARIMA) model. Functional methods outperform the conventional ARIMA model in both short and medium-term forecast horizons. In addition, performance of functional methods in forecasting beyond one hour is also found to be consistent. Figure 1.2 illustrates the research overview.

2. ON TRAFFIC FLOW PATTERN SHAPE CLASSIFICATION AND ANALYSIS

2.1. INTRODUCTION

The advancement in data acquisition has resulted in better insights into traffic dynamics. Traffic researchers have been using techniques from contemporary fields to extract meaningful information from available data. One such technique known as clustering. The primary objective of clustering is to identify relatively mutually exclusive, homogenous groups within a sample of entities based on the similarities present between the individual entities. The application of clustering techniques is therefore a natural one, when there is an interest in determining the presence of, identifying, or recognizing patterns of interest within a historical traffic dataset. Furthermore, clustering is often an exploratory step prior to the development of any model. Specifically, data analysis / mining is done prior to the development of any traffic related paradigm.

In the literature a variety of approaches to analyze traffic patterns for different traffic dimensions: congestion recognition, traffic accident recognition and general traffic pattern recognition have been presented. Specifically, in Zhu and Barth (2006), a study of vehicle activity patterns to classify different congestion levels through wavelet analysis combined with principal component analysis is presented. Similarly, Lozano (2009) present a study of congestion levels at intersections and classifies congestion through the application of K-means clustering. In addition, a new framework toward a real-time automated recognition of traffic accident based on Histogram of Flow Gradient (HFG) and statistical logistic regression is presented in Sadek et al (2010).

Due to the insight gained through clustering and pattern analysis of historical traffic data, numerous studies have been undertaken where the proper grouping of data is the primary goal. For example, Rakha and Van (1995) studied flow and speed variations between days and ascertained on the basis of an Analysis of Variance (ANOVA) that traffic flows on a freeway vary significantly within a week between two groups, in which {Tuesdays, Wednesdays,Thursdays} define a group and {Mondays, Fridays, Saturdays, Sundays} define another distinct group. However, the days were predefined into core week and weekend days prior to analysis. In Wild (1997), daily traffic is classified into six distinct groups, in which days are discriminately placed into groups based on daily flow characteristics. In the clustering of traffic patterns there is a lack of consistency regarding the classification of days into a clear set or number of groups. In an effort to clearly classify traffic patterns, Chung (2003) classified daily travel time curves from the Tokyo Metropolitan Expressway and concluded that AM and PM periods result in distinct groups. Specifically, Chung (2003) states that grouping of the AM period consisted of weekdays, Saturdays and holidays (including Sundays). He continues, by stating that the PM period, however, should be treated separately and implies that the grouping of PM period through regular clustering methods yields weak grouping. The results of both of these studies are difficult to interpret and summarize. This difficulty is due in part to the customized clustering techniques used, which Soriguera (2012) suggests are not suitable for traffic patterns due to the stochastic nature of travel time data. In Chung (2001), agglomerative hierarchal clustering of daily traffic flow data was completed in an effort to develop an effective short-term prediction model. The authors conclude that their model seeks to capture the distinctive shape of the traffic flow profile through the use of a historical average, however, the authors state that the model is highly sensitive to outliers in the data. Outliers can be caused by a malfunctioning sensor, weather or other external factors. Perhaps the importance of proper clustering is best stated by Rakha and Van (1995), which states that anomalies in the historical data can cause models to deviate or become ineffective. Examples of such models, numerous time series based models, which under perform in predicting the time horizon, refer to Nicholson and Swann (1974); Hogberg (1976); Ahmed and Cook (1979); Ahmed (1983); Okutani and Stephanedes (1984); Stephanedes et al (1990). The model performance results from the historical dataset that was used to parameterize the model. Hence, through proper grouping/ clustering better models can be parametrized for specific situations of interest. Research suggests that pre-definition is necessary and there appears to be a consensus that daily traffic patterns can be defined into four basic groups namely the Monday through Thursday, Fridays and Days prior to holidays, Saturdays except holidays and Sundays with Holidays groups. Recently, Soriguera (2012) proposed a hierarchical multi-step clustering technique for traffic pattern classification. The multi-step clustering procedure consists of separating standard from non-standard days in its first step. The second step consists of seasonal classification. The third step seeks to separate days, which were not grouped in the previous two steps. All these works suggest that shape is important, yet none of them focus on the development of classification based on shape or include shape based methodologies in their classification techniques.

A common convention of all previously mentioned research efforts is the lack of classification of traffic patterns solely based on their shape characteristics, or more specifically, exploit the information in the shape of the traffic pattern to the fullest, which to the best knowledge of the author has not been fully carried out with a high degree of detail or through the allocation of shaped based methodologies. To this end, the research presented herein seeks to perform an analysis of classification of daily traffic profiles into distinct groups based on their inherent shape. The shape characteristics of a daily profile is a direct result of traffic behavior. An example of this is the bi-modal shape found on a common weekday caused by the AM and PM peak periods, which are commonly known as morning and evening rush hours respectively. Other behaviors are sharp dips observed during a peak as it is often related to the occurrence of an accident as the flow reduces to a crawl or suggest that congestion has reached high enough levels to be considered a ”traffic jam.”

This study seeks to fill the gaps prevailing in the literature by introducing a shape based clustering technique, which does not require any pre-classification of existing patterns in the data. The advantage of the proposed methodology eliminates bias of heuristic preprocessing such as those found in Wild (1997); Chrobok (2004) . In addition, the methodology proposed reduces the computational cost (due to smaller number of observations required to describe a daily profile) and it is robust to fluctuations in flows where shape is preserved. It is effective in capturing outliers into separate groups, which provides a strong basis for precise prediction modeling and relevant research effort within traffic contexts. Above all, traffic profiles are clustered based solely on their shapes and therefore making trivial the requirement of analyzing different flow characteristics: total flow, peak flow, peak period and off peak time separately.

In Weijermars and Van Berkum (2005); Weijermars et al (2007) clustering is employed to analyze traffic flow patterns. The study concluded that shape of the daily flow profiles may differ between different types of days and motivates the need for such an analysis of traffic pattern shape characteristics. Through proper classification of historical traffic data the performance of such models can be improved as proper classification leads to better parameterizations of models. In this study, we merely present a classification

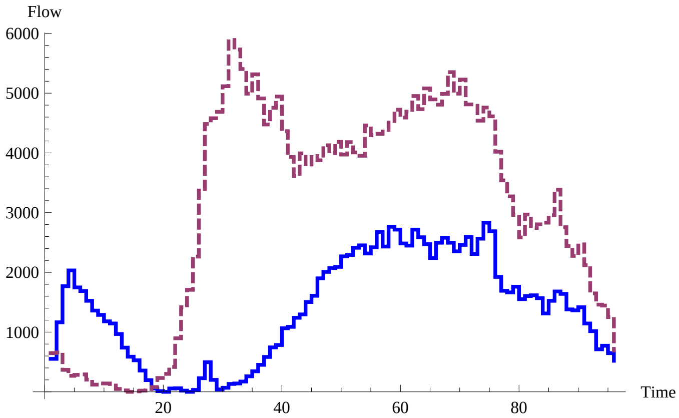


Figure 2.1. Two daily traffic flow profiles for Thursdays Jan. 1, 2004 (bold) and Jan. 22, 2004 (dashed) are given, which demonstrate unique shapes. Time is in 15 minute interval.

technique based on shape. To illustrate this point, consider Figure 2.1, which illustrates two Thursdays in the same month, however, their shapes differ greatly. In this case Jan. 1, 2004, which is a common holiday, differs greatly to the other Thursday that inhibits regular expected shape characteristics where two peaks are observed corresponding to morning and evening rush hours. Hence, classification of these days according to traditional clustering methods would classify these two days into different groups. To this end, this research seeks to determine if shape can result in groups of similar daily behavior.

The contribution of the research presented herein is the use of Mathematical Morphology (MM) tools to achieve proper classification of daily traffic profiles. Specifically, in this study the use of the Granulometric Size Distribution (GSD) is emphasized. The GSD is developed by employing commonly used MM tools for image processing. Refer to Serra (1982) for a seminal book on the topic. Through the employment of such tools the development of an analysis of shape is carried out in an effort to generate interpretable classification of historical daily traffic profiles. An additional advantage of these tools is that it considers the entire daily profile as a single datum. Hence, the development of unique granulometries for each day are contracted and are used to cluster days into distinct groups. The use of Partition Around Medoids (PAM) algorithm or otherwise referred to as the K-Medoids algorithm is employed to carry out the classification of daily profiles.

The remainder of this study is organized as follows. In Section *Mathematical Morphology*, the MM tools used are detailed regarding their employment to the analysis of signals rather than images. This is followed by the description of the methodology undertaken to cluster historical daily profiles based on their shape characteristics in Section *Methodology*. A brief description of the data used to validate this new proposed application of MM tools is given in Section *Data Details*. Lastly, the results/discussion and conclusion are given in Sections *Results* and *Conclusions* respectively.

2.2. MATHEMATICAL MORPHOLOGY

Mathematical Morphology is a theory that provides a number of useful tools for image analysis, which is based on the assumption that images consists of structures that may be handled by set theory. In addition, the tools and methods that MM provides can be naturally applied to the analysis of signals. In Gaston-Romeo et al (2011), MM tools were applied to signals in an effort to analyze daily solar radiation time series curves. Similarly, in Guardiola and Mallor (2013) unintended electromagnetic emissions from wireless communication devices are analyzed in the frequency domain. In both previously mentioned works, curves are considered to be bi-dimensional images on which morphological operators are applied to gain information regarding the inherent shape of the two phenomenons.

2.2.1. Opening, Erosion, and Dilation Operators. In this study, three MM operators are used to study the shape of the daily traffic flow profile. Specifically, Dilation, Erosion and Opening operators. These operators are used to construct a daily traffic profile's *Granulometric Size Distribution* (GSD). Consider an entire daily time series traffic flow profile represented by function $f(t)$ which it takes only positive values (e.g. there can be no negative traffic), $f(t) \geq 0$. The MM operators extract the shape of the structure by probing it by a known shape called a Structuring Element (SE). Theoretically, a subgraph of the original function, $f(t)$ is defined as $SG(f(t)) = \{(t, y) : 0 \leq y \leq f(t)\}$. For better assimilation, a vector with small flow values is used.

1: Erosion of a function– $f(t)$ by a $SE\ B[-1, 1]$ is the function τ_B defined as $[\tau_B f(t)] = \min_{b \in B} f(t + b)$.

2: Dilation of a function– $f(t)$ by a $SE\ B[-1, 1]$ is the function μ_B defined as $[\mu_B f(t)] = \max_{b \in B} f(t + b)$.

3: Opening of a function– The combination of erosion and dilation operation is called opening. The erosion firstly shrinks the image followed by dilation which expands it thus combine together to explain opening operation. Mathematically expressed as $\gamma_B f(t)$= $\mu_{\breve{\mathbf{B}}}\{\tau_B f(t)\}$= $\mu_{\breve{\mathbf{B}}}\tau_B f(t)$. Where $\breve{\mathbf{B}} = \{-b : b \in B\}$ is the symmetric set with respect to the origin of B. In the case of a symmetric structuring element, B, then $\breve{\mathbf{B}} = B$. In our case structuring element is a unit square.

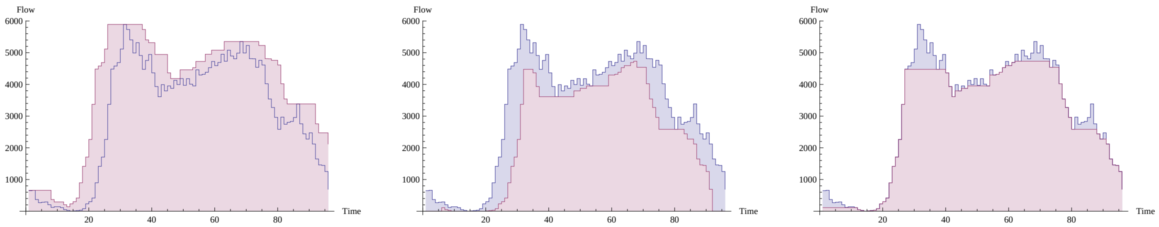
Figures 2.2a, 2.2b, and 2.2c illustrate the MM operators of *Dilation*, *Erosion* and *Opening* respectively applied to the Jan. 22, 2004 daily flow profile. The construction of the corresponding GSD function is the successive application of the *Opening* operator with an SE increasing in size.

4: Granulometric Size Distribution– The GSD, $F_{f(t)}(\beta)$ can be defined as

$$F_{f(t)}(\beta) = 1 - \frac{A(\beta)}{A(0)}. \tag{1}$$

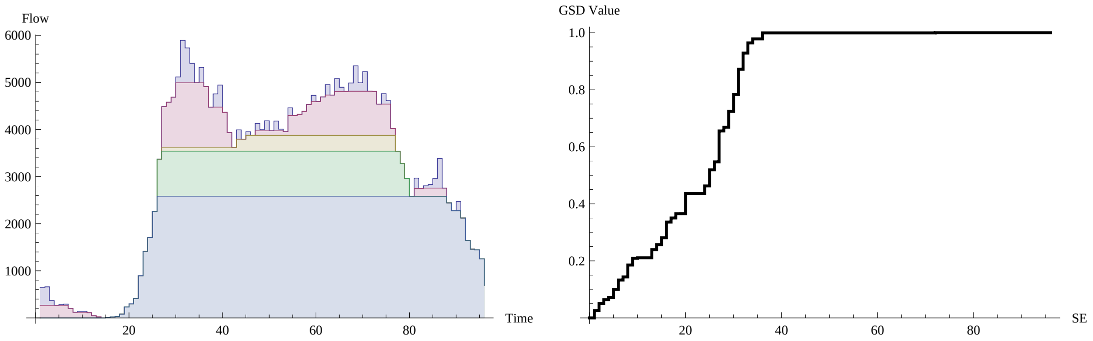
Where, $A(\beta) = \int_S \gamma_B(f)(t)dt$, which defines the area of the curve under the opening of the function. Hence, $A(\beta)$ is the area under the opening of the function when the structuring element, SE, is of size β and $A(0)$ is the area of the original traffic profile. The GSD is constructed through the successive application of Equation 1 while increasing the size of the SE. Consider Figure 2.3a, where a set of $SE = \{2, 15, 25, 30\}$ is applied to the Jan. 22, 2004 daily traffic profile and is depicted as layers. In Figure 2.3b the corresponding GSD for Jan. 22, 2004 is presented by applying $SE = \{0, 1, 2, \ldots 96\}$. Note that 96 is a result of taking the 1,440 minutes in a day in 15 minute intervals resulting in 96 total intervals. Thus, the SE is increased from 0 to 96 in order to determine the shape of the traffic profile curve.

Through the use of the GSD, difference in shape is emphasized. Thus, a comparison of the two Thursdays' (e.g. January 1 and 22 of 2004) GSDs depicted in Figure 2.4. This illustrates that the two profiles have distinct GSDs.



(a) Dilation operator (b) Erosion operator (c) Opening operator

Figure 2.2. The application of the MM operators on the Jan 22, 2004 daily flow profile using an SE of size $\beta = 5$.



(a) Successive application of the opening. (b) Corresponding GSD for Jan. 22, 2004.

Figure 2.3. Construction of the GSD of traffic flow profiles.

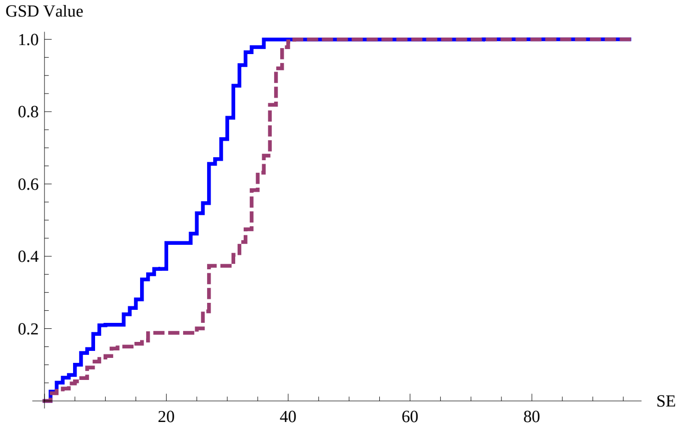


Figure 2.4. Comparison of GSD of daily profiles illustrated in Figure 2.1.

2.3. METHODOLOGY

Following analyses is carried out to highlight the use of mathematical morphology. First, a robust study is carried out to demonstrate that the proposed use of the GSD function can effectively deal with various data situations. Secondly, a classification study is carried out and analyzed for the large data set spanning from 2004 through 2008.

2.3.1. Analysis of Robustness. In this study the GSD's flexibility is highlighted. Specifically, we use a single daily profile of January 22th, 2004 to highlight how the GSD function behaves when a daily profile is scaled and shifted.

First, scaling the daily traffic profile is accomplished by multiplying the daily profile with a set scalars. Specifically, the set $\Gamma_i = \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, 2, 3, 4\}$ for $i = 1, 2, \ldots 7$, which generates seven unique curves that have the same shape. Let $f(t)$ denote the daily profile then scaling it by a scalar Γ_i can be denoted as $\Gamma_i f(t)$. Each of the scaled profiles is used to generate their respective GSD functions. Figure 2.6a illustrates these unique set of curves with the profile 4 being the original profile as it is scaled by 1. An SE of size six is used to generate the corresponding GSD curves for each these profiles. Hence, the set of $SE(\beta)$, where $\beta = \{6, 12, 18, \ldots, 96\}$ is used. The aim is to demonstrate that simple scaling of the daily traffic behavior will result in the same GSD function. Hence, if traffic behaves the same throughout a given day the corresponding GSD function will not be influenced as the resultant analysis of the profile's behavior (e.g. shape) has remain unchanged.

Secondly, the shift of the daily profile by a scalar is accomplished in a similar fashion to scaling of the profile, however, the daily profile $f(t)$ is shifted upward and can be denoted as $f(t) + \Phi$, where Φ is a simple increase of the profile's base area. The scaled set is accomplished by shifting the daily profile of Jan. 22th, 2004 by a set of scalars Φ_i. Specifically, the set $\Phi_i = \{0, 1, 2, 3, 4\}$, for $i = 1, 2, 3, 4$ is used. Thus, each profile will be shifted by applying the following equation, $f(t) + \Phi_i * 500$, which for example when $\Phi_i = 2$ will result in a shift of 1000. This results in the shift of the daily traffic profile, which is illustrated in Figure 2.7a where trace 1 is the original profile as a shift of 0 is applied. The resulting set of shifted profiles are used to generate their respective GSD function. Similarly, to the scaling study an SE of size six is used to generate the corresponding GSD functions of the shifted profiles. Hence, the set β is again applied. The goal of the this study is to demonstrate that the GSD is an analysis of the shape through area. Hence, a shift of this type results in a baseline flow area underneath the entire daily flow profile curve. The GSD will increase as the shift is increased which allows for the detection in growth of traffic volume. As there is a large area that is common to the entire profile the shift should be more easily detectable when the SE size is large. Thus, the shift should result in large increase in the corresponding GSD values for whose size, β, is largest.

Through both of these studies we highlight that the GSD is robust to change of traffic behavior that results in unique shapes. The scaled study aims to demonstrate that the amplification of the shape results in the same GSD, while the shift corresponds to an analysis of the growth in volume. The results and discussions of this study are highlighted in section *GSD Robustness*.

2.3.2. Classification Study. This study aims to highlight the use of the morphological tools and methods as means to classify traffic behavior based on shape characteristics. First, a daily traffic flow profile is converted to their corresponding GSD curve. This is done for all 1,827 days within the chosen data set. The data details are given in Section 4.4.1. Secondly, clustering is conducted on the GSD curves. Specifically, clustering is done through the use of the Partition Around Medoids (PAM) or K-Medoids method. Specifics regarding this clustering algorithm are summarized in Appendix 6. The clustering involves the use of the *Correlation Distance*, which is defined as the distance,

$$D_c(u, v) = 1 - \frac{(u - \bar{u}).(v - \bar{v})}{(|u - \bar{u}|.|v - \bar{v}|)},$$

and gives the correlation coefficient distance between vectors u and v. Where $\bar{u}$ and $\bar{v}$ are the mean of the vectors respectively. Gap statistics is used as method of optimization for clustering. The *"Gap"* test compares dispersion of the clusters generated from the data derived from a sample of null hypothesis sets. The null hypothesis sets are uniformly randomly distributed data in the box defined by the principal components of the data. This distance is chosen as we want to assure that the traffic profiles are clustered according to their relationship with one another based on the GSD curve generated from their shape. In this way similar patterns are grouped together, where patterns with low relationship to other curves are clustered together. In this manner, curves that inhibit similar characteristics will dictate a group. In addition, a low tolerance value is chosen to be 2. Typically, the larger values of tolerance favors fewer clusters and it is common to use values within the range of 1-5. The tolerance sets the sensitivity. Lastly, hierarchal clustering is performed on groups that are loose (e.g. exhibit large variety). In this step, loose groups are further analyzed. For this step, a hierarchal agglomerate method is used in combination with the previously mentioned application of the PAM with the ”gap” statistic. The reason behind the application of two different clustering methods is to assure consistency as well as to get more insight into looseness of cluster. Furthermore, the agglomerate method is preferred over other clustering techniques is due to feature of dendrogram, which is a illustration of connections of lines that represent the fusion of clusters, and lengths represent the degree of dissimilarity between clusters. This allows to gain insight into within cluster organization e.g which curves within the clusters are closer to each other. This analysis is performed on groups of high looseness to determine type of shapes contained by the cluster. In the traffic context, this could be curves that are misclassified in previous analysis or have some sort of abnormal shape. Through the employment of this methodology a new clustering of traffic curves is accomplished. The results of this methodology are explained in detail in the proceeding section of this study. Specifically, refer to section *Cluster Analysis* for the primary clustering of all 1,827 days. Section *Analysis of Group 7* contain the results of the exploratory hierarchical clustering of groups with loose clustering.

2.3.3. Shape Change Analysis . The goal of this study is to highlight GSDs capability to cluster sub shapes correctly. The GSD is directly associated to the shape of the profile and not time. Hence, it is highly sensitive to changes in the shape of the daily profile regardless of when those changes occur in time. In other words, a profile and its mirror image with respect to time will generate the same GSD. In an effort to investigate the GSD's capability to analyze changes in the overall shape by analyzing sub shapes present in a profile. Consider that each daily profile is segmented into two segments (e.g. am and pm), where one segment is between 12 am (midnight) and 11:59 am and the other between 12 noon and 11:59 pm, creating two sub shapes. Next, two GSDs are constructed for each am and pm segment respectively. These two GSD vectors are then combined and clustering is performed. However, for a random set of the combined vectors the am period GSD is switched with the pm period and added to the the data set. This study uses all Tuesdays for the year 2004 from January 1, 2004 to December 31, 2004. This results in 52 original profiles where a random set of 6 were chosen to have their am and pm period GSD switched. They are then clustered with the same days with un-switched patterns. Clustering is performed with 52+6=58 profiles by invoking the *FindClusters* function in $Mathematica^{©}9$

for the PAM algorithm with the optimization of groups through the ”gap” statistic and a correlation distance. The same clustering as described in the preceding section *Classification Study*. The results of this study is highlighted in section *Switched Subshapes*.

2.4. DATA DETAILS

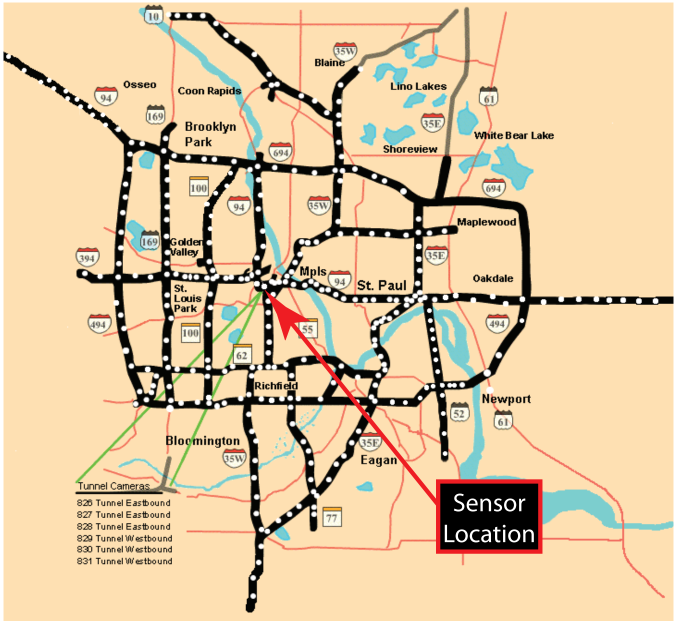


Figure 2.5. Map of Minneapolis MN U.S.A Highway and Freeway system. Arrow indicates position of sensors on I-94.

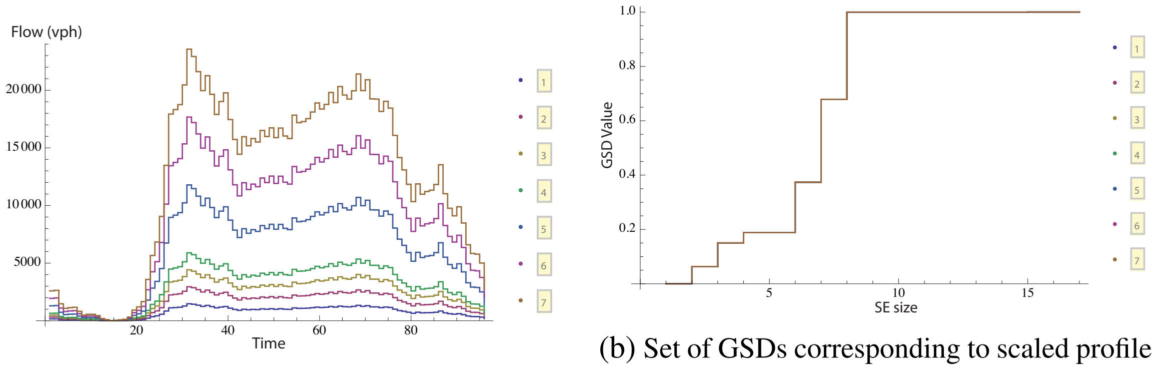
The study herein is based on traffic data collected from the I-94 within the Twin Cities Metro area, Minnesota. Data is obtained at station S110 I-94 East Bound/T.H. 65 which has 3 loop detectors D497 94/TH65E1, D498 94/TH65E2 and D499 94/TH65E3. I-94 has 3 lanes in each direction and the station provides composite detector data from three detectors in the eastbound direction (Refer to Figure 2.5). The analysis period is from 1st January 2004 to 31st of December 2008. The detectors measured and logged the flow for each of the three lanes at 30 seconds intervals. For this study, data is aggregated over 15 minute intervals (96 points per day). Weijermars and Van Berkum (2005) found that 15 minutes data produce better results as the fine grain variations removed. At the selected location, traffic data of 5 years translate into 1,827 days and 175,392 observations. Simply, there are 261 of each of the week within the data set for all 5 years.

2.5. RESULTS

In this section the results pertaining to the proposed analysis detailed in Section *Methodology* are presented. The implementation of the analysis is performed within the $Mathematica^{©}9$ software package using the built in functions of Mathematical Morphology. Specifically, the GSD is constructed using the *Opening* function. Clustering is achieved through the implementation of the *FindClusters* function. Both of these functions are well tested and documented.

2.5.1. GSD Robustness. The results in this section highlight the robustness of GSD function with regards to shape scaling and shifting. First, the scaling of a traffic profile occurs in the traffic context when traffic behavior remains the same but the magnitude of the flow is increased or decreased. The Jan. 22th, 2004 day is a Thursday and contains characteristics common to a weekday traffic profile. These characteristics are two peaks, which correspond to the morning and evening rush hour behavior. The goal of this study is to demonstrate that the GSD, which is derived through the morphological operators, analyzes the shape not the magnitude of the shape. Figure 2.6a illustrates the scaling of the Jan. 22th, 2004 profile. The GSDs for each of the scaled versions of the daily traffic profile are illustrated in Figure 2.6b. It is easy to notice that the corresponding GSDs for each of the scaled profiles results in an identical GSD. For example, when the SE size is 36, the GSD value is $GSD(36) = 0.373648$ is identical to all 7 modified profiles. The importance of this simple analysis is that the proposed method seeks to classify traffic based on their shape pattern that in turn describes specific traffic behavior. The GSD is highly robust to scaling effects and therefore its use as a means to classify days based on similarity of pattern is natural and appropriate.

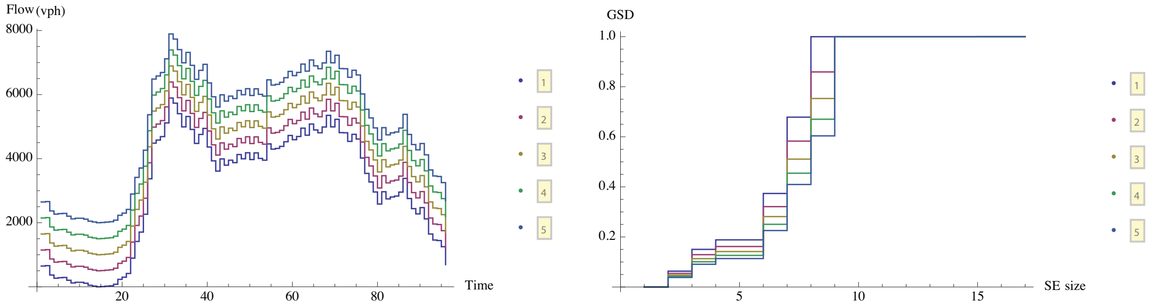
Next, we analyze the shift of the traffic profile, which are illustrated in Figure 2.7a. The Jan. 22th, 2004 profile is shifted upward with an ever increasing baseline traffic. This baseline is the area underneath the entire function (see Figure2.8). This results in a GSD that will have larger values for larger SE sizes. This is due to a large percentage of the shape of the profile is now dictated by this baseline or shift of the traffic profile. In Figure 2.7b, this is clearly illustrated as the shift is increased, the GSD's value at larger SE sizes creates a more predominant difference between each corresponding GSD. This highlights GSD sensitivity , as days with similar baseline traffic will be clustered together in combination with profiles that have similar shapes. This will occurs due to the distance



(a) Scaled Jan. 22th, 2004 by set of scalars A.

(b) Set of GSDs corresponding to scaled profile set.

Figure 2.6. Robustness of the GSD function scaling.



(a) Shifted Jan. 22th, 2004 by set of scalars in B.

(b) Set of GSDs corresponding to shifted profile set.

Figure 2.7. Robustness of the GSD function shifts.

measure increasing between shapes of similar patterns but larger shifted flows. Thus, the GSD method results in the classification based on behavior, which is dictated by the shape, and volume as the GSD is sensitive to both. That is to say that daily profiles that have similar shape and volume will be grouped together as the baseline area underneath the traffic profile will increase the dissimilarity among profiles, but daily profiles that have similar shapes except for an upward or downward shift due to the addition of a constant value to daily traffic flow will be less likely to be grouped together.

2.5.2. Cluster Analysis. In this section the results pertaining to the cluster analysis detailed in section *Classification Sudy* is presented. The daily traffic flow profiles broadly exhibit a variety of shapes. The dominant shapes are the well known unimodal and bimodal

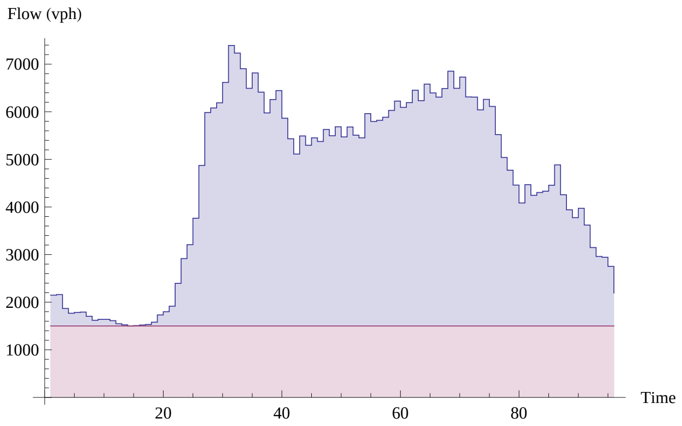
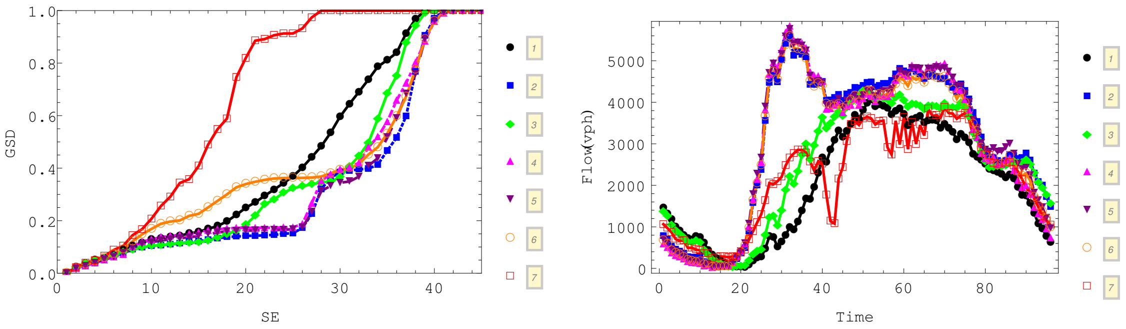


Figure 2.8. Shifted profile with *opening* performed with an SE size of 42.

shapes. The unimodal shape corresponds to the single peak relating to the maximum flow intensity and is typically referred as non-working day behavior. The bimodal flow shape corresponds to morning and evening peaks, which is commonly referred to as working day behavior. However, we find that while most of the days fit into these two dominant shapes, some cluster groups inhibit a variety of different shapes.

In Figure 2.9, the medians of each group's GSD (Figure 2.9a) and original traffic flow profile (Figure 2.9b) illustrate that the clusters have small but detectable differences in shape. Specifically, it easy to observe that the medians of groups two, four, five, and six have similar shapes, however, the four groups appears to have small variation in the width and magnitude of the two peaks. Conversely, groups one and three seem to have the unimodal shape with major differences in the location of the peak, as well as differences in its magnitude and duration. Lastly, group seven illustrates that this group contains highly variable shapes that differ from the commonly accepted norm. This can be seen as the group's median both in the GSD and original traffic flow profile contains many peaks and valleys.

The composition of the seven groups are provided in Table 2.1 as the percentage of type of day. Three types commonly used are defined, which are *business days, holidays* and *weekend days*. A *business day* is defined to be a regular working week day excluding holidays. Similarly, *weekend* days are regular weekend excluding holidays. It can be seen that that groups two, four, five and six are comprised primarily of *business days*. Groups



(a) GSDs

(b) Original flow

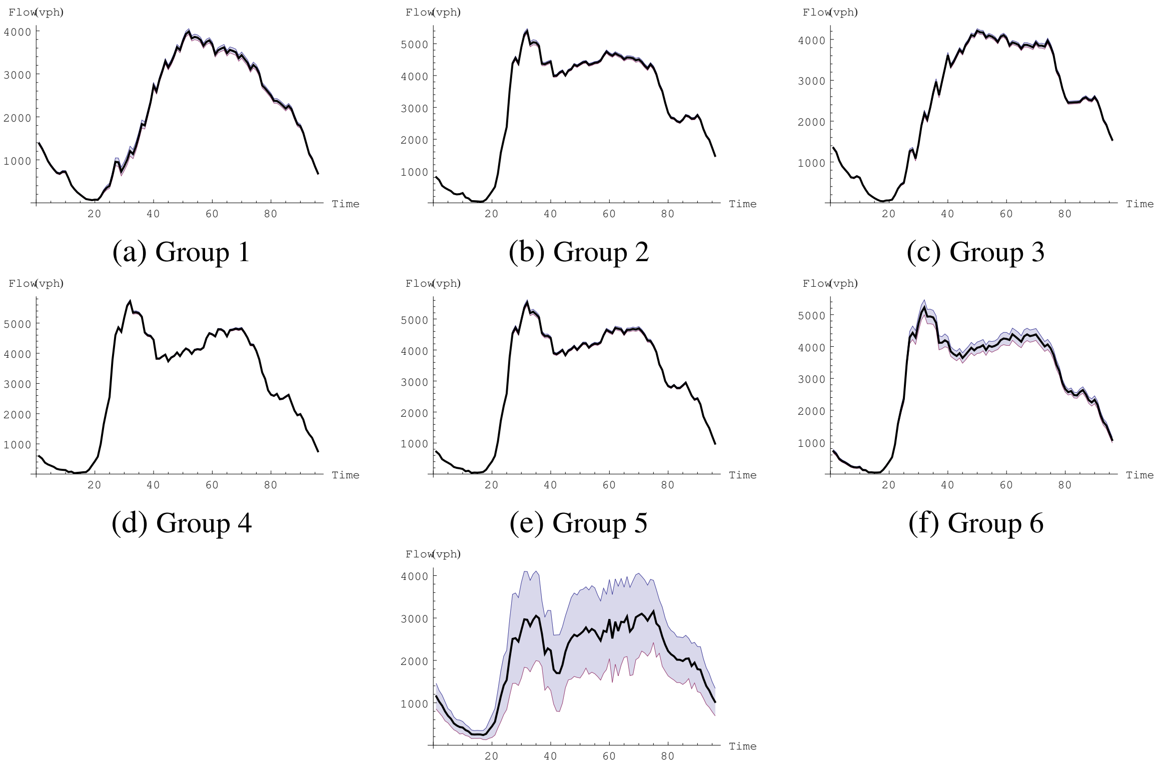
Figure 2.9. Corresponding median GSD (2.9a) and corresponding traffic flow profile medians (2.9b) for each of the seven groups.

Table 2.1. Group composition

Group	Business Days	Holidays	Weekends	Total Days
1	5.5%	8.12%	86.4%	308
2	99.6%	0.4%	0.0%	251
3	2.0%	2.0%	95.9%	245
4	97.1%	2.2%	0.7%	547
5	98.2%	0.23%	1.6%	385
6	89.3%	6.7%	4%	75
7	50%	0.0%	50%	16

one and three are primarily comprised of *Weekend* days. Lastly, group seven contains equally business days and weekends. This is interesting as there must be a shape that drives the composition of that group.

In an effort to visualize the inherent differences as well as the tightness of the groups the 95% Confidence Interval (CI) is plotted around the mean curve of each group. It is important to note that since clustering is used to create these groups, the standard deviation here refer to the looseness of a cluster at a given time rather than the variance of a population. The variation of the population should be computed by all clusters. By the very nature they were created, their standard deviations are going to be small and the confidence intervals will not overlap. Specifically, the CI is calculated as the $\bar{x_i}(t) \pm 1.96\frac{\sigma_i(t)}{\sqrt{n_i}}$, where $\bar{x_i}(t)$ is the mean of group i at time t, $\sigma_i(t)$ is the standard deviation at that time and n_i



(a) Group 1

(b) Group 2

(c) Group 3

(d) Group 4

(e) Group 5

(f) Group 6

(g) Group 7

Figure 2.10. Group 95% confidence intervals

is the number of curves in group i. These mean curves and their corresponding CIs are illustrated in Figure 3.4. It can be seen that the 95% CIs of group one, two, three, four and five are very tight. Hence, this validates that the grouping is effective and tight as similar shapes and behavior have been clustered together. For example, Table 2.1 shows that 99.6% of group two is comprised of regular business days. These results illustrate that the combination of groups two, four, five and six is possible as they inhibit similar shapes and flow volume characteristics. However, Figures 2.10f and 2.10g display that these groups have looser grouping than the other groups. Thus, by invoking hierarchal clustering within these two groups more insight into the composition of these two groups is possible. This analysis is carried out in the preceding sections of this study (refer to section *Investigating Missing data and Special days*).

The regular patterns mainly working days (bimodal) and non working days (unimodal) are well documented in literature. However, the shape analysis has resulted in the creation of groups that inhibit small differences in the peak characteristics and late evening and early morning period behavior. The focus of the remaining study will be on the groups having non regular shapes.

2.5.3. Investigating Missing Data and Special Days. Groups six and seven contain different shapes from the routine traffic patterns observed in other groups. In addition, these group's standard deviations shows that the grouping is looser than the other groups. This motivates further analysis into these groups. Possible reasons for the shapes observed in these two groups are but not limited to: incidents, missing data from malfunctioning of detectors, weather conditions, congested traffic conditions due to bottle necks and various social events could result in unique shapes.

Each of these types have different characteristics and impacts. For example, incidents produce different conditions depending on the location of the incident; upstream or downstream Karim and Adeli ((2002)). Different patterns can be formed on the upstream or the down stream side of capacity-reducing incidents. This can be observed as a valley in the traffic flow profile during peak periods as traffic speed and flow decrease due to partial lanes closure, which create bottleneck conditions. On the other hand, flow pattern changes are more pronounced in the case of upstream sensor location to an incident. The malfunctioning of the loop detector is a common issue (e.g giving zero values) and have impact on data mining techniques. Extreme weather conditions like snowfall and low visibility reduces speed, which can be observed as decrease in magnitude in flow profiles. As the freeway in question is an urban freeway, a social event in the near vicinity is likely to effect the flow patterns as well.

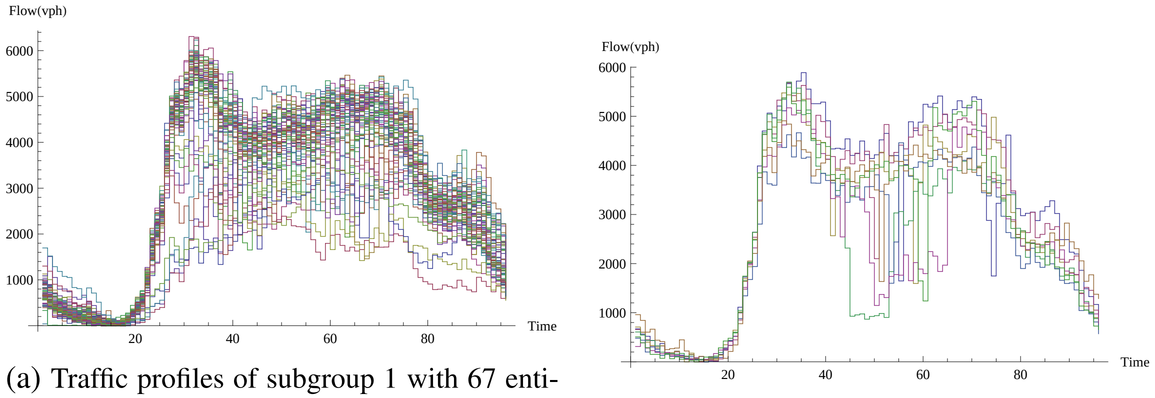
In order to gain an insight into the identification of such events through their corresponding daily traffic profile shape. Hierarchical clustering was conducted on group six and seven GSD's to gain insight into the composition of these loose groups. Sub clustering is performed to further decompose possible data or traffic behaviors. Specifically, clustering within group six and seven was carried out using two techniques, which were Agglomerate and the PAM, using the same *correlation distance*. The only difference is the Agglomerate technique requires the identification of linkage, which is defined as Unweighted Pair Group Method with Arithmetic Mean (UPGMA), $UPGMA = \frac{1}{|F||E|}\sum_{f\in F}\sum_{e\in E} d(f,e)$, where d is distance and UPGMA is the average linkage between two set of observations F and E.

2.5.4. Analysis of Group 6. Clustering is performed on the original Group 6 through both the PAM and Agglomerative methods. The results are interesting as both methods resulted in two subgroups. However, the composition of both groups differ. Table 2.2 contains information about the composition of the subgroups. First, it is interesting that through the agglomerative clustering one dominant group is found containing the majority of the observations whilst the other only contains 8 observations of the total 75 (refer to Figure 2.11). The subgroup containing 8 has a very interesting pattern as all the profile have a major fluctuations where the flow decreased to a relative low level for an extended period of time. This behavior can be related to incidents or construction/maintenance activities where the traffic was reduced to a bottleneck and flow resumed but at low levels. The other subgroup contains a variety of observations, however, upon close inspection of this group it is determined that this group consists of profiles that have major fluctuations through the day. Specifically, this group contains large fluctuations in their profiles with immediate recovery (e.g. the profiles contain dips in their profiles during peak time periods). In addition, apart from a single observation no zero entries are observed in this subgroup. It suggests that any major change located in tail region of the profile may not have a significant impact on classification. The methodology proved effective in clearly differentiating between abnormal days from incident days within the group. The PAM method similarly clustered the groups, however, more evenly between its two subgroups (refer to Figure 2.12). The PAM method mixed both long-term fluctuations with low flow level days. It is more difficult to present a similar argument for the PAM method results. However, some common characteristics is that one subgroup again is primarily dominated with days which contain fluctuations, while the other subgroup contain low peaks, low flow levels, and small fluctuations that have quick recovery spanning 15-30 minutes.

Different types of non recurrent patterns can be seen in Figure 2.12b. Three patterns: abnormally high sustained evening peak and high late night peak representing some social activities and abnormal sustained low flow representing extreme weather conditions. The change in flow patterns due to incidents are identified with respect to incident location in Figure 2.12b.

Table 2.2. Subgroup composition within Group 6.

Method	PAM		Agglomerative	
	Group 1	Group 2	Group 1	Group 2
Number of Days	22	53	67	8

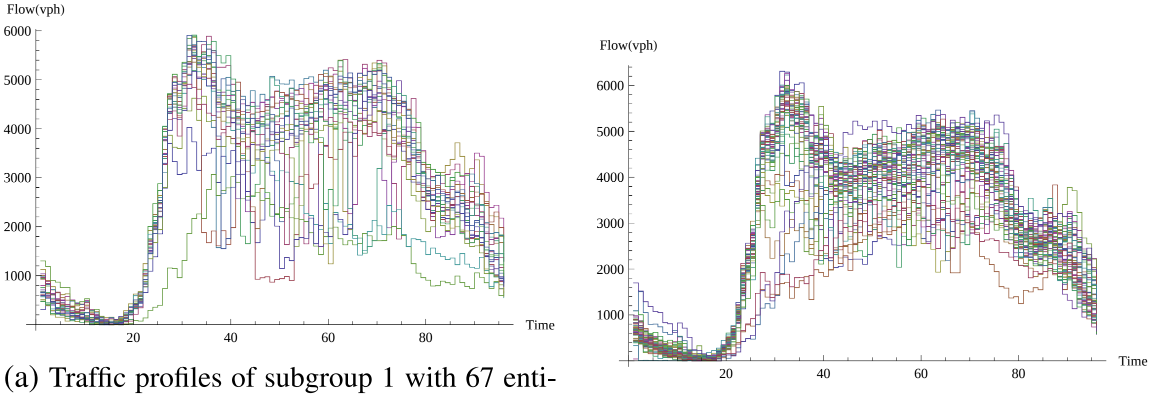


(a) Traffic profiles of subgroup 1 with 67 entities.

(b) Traffic profiles of subgroup with 8 entities.

Figure 2.11. Two subgroups through the agglomerative clustering

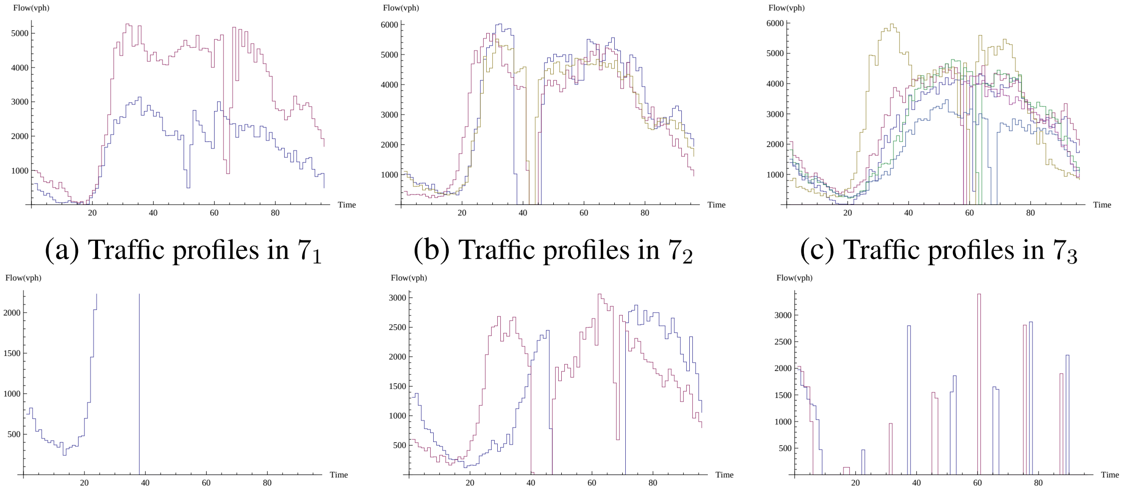
2.5.5. Analysis of Group 7. Group seven inhibits a great deal of information. The results of further analysis into this group yields motivation for shape based analysis of traffic profiles. First, similarly to the study performed on Group six both the Agglomerative and PAM methods are performed on Group seven. The PAM results in six subgroups where the Agglomerative method results in four subgroups. Due to the large number of subgroups for the sake of clarity the subgroups will be referred to as 7_i, which is the subgroup i for $i = 1, 2, \dots 6$ of the original group seven. The six subgroups contain $7_i = \{2, 3, 6, 1, 2, 2\}$, which is the number of traffic profiles within each subgroup. Thus, 7_1 contains only two observations. The interesting result is that the last three groups (e.g. 7_4, 7_5 and 7_6) are identical under both methods. These three subgroups contain the exact same observations and are illustrated in Figures 2.13d, 2.13e, and 2.13f. Similarly, the first three subgroups of



(a) Traffic profiles of subgroup 1 with 67 entities.

(b) Traffic profiles of subgroup with 8 entities.

Figure 2.12. Two subgroups through the PAM clustering



(a) Traffic profiles in 7_1

(b) Traffic profiles in 7_2

(c) Traffic profiles in 7_3

(d) Traffic profiles in 7_4

(e) Traffic profiles in 7_5

(f) Traffic profiles in 7_6

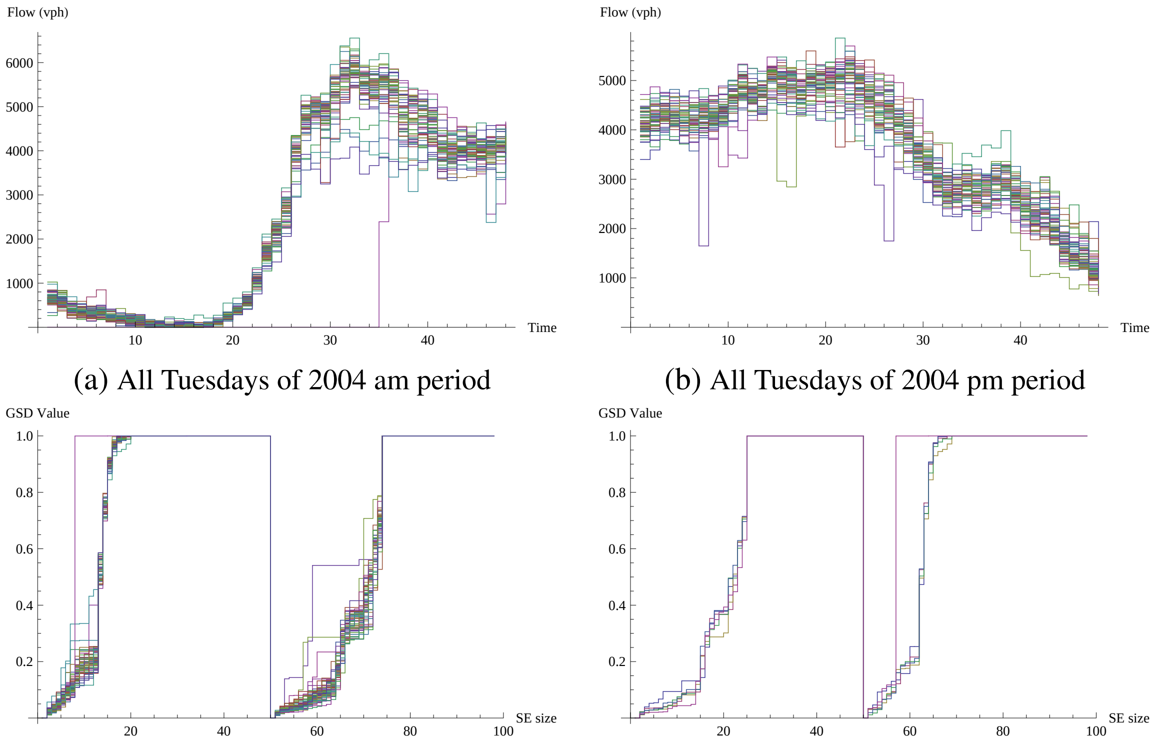
Figure 2.13. PAM subgroups of Group Seven, 7_i, for $i = 1, 2, \ldots 6$.

PAM (e.g. 7_1, 7_2 and 7_4) correspond to the first group of the agglomerative method. Hence, the union of these three subgroups is the corresponding observations found in subgroup 1 of the agglomerative method.

In terms of traffic this decomposition is the most interesting found. It clearly demonstrates the GSD's capability to analyze shapes. Specifically, 7_1 contains two observations where the general shape is conserved, yet, they both have an incident type characteristic related to congestion. Conversely, 7_2, 7_3 and 7_5 appear to have characteristics that have been noted in literature as behavior relating or caused due to a road closure, maintenance activity or malfunctioning detectors. Each of these three subgroups contain profiles where there is a significant span where the detector has collected no data. Furthermore, each group differs in the amount of time. The interesting aspect is that there is traffic prior and after the 0s. In 7_4 there is only one observation and it must be caused due to a detector malfunction as after a certain time there is no flow for the remainder of the day. Similarly, 7_6 the detector seems to be turning on and off throughout the day only collecting for small segments of time.

The ability of the proposed methodology is in part the capability to differentiate between shapes. Specifically, the capability of the GSD detecting the abnormal, missing and/or non-realistic loop detector data. This analysis illustrates the GSD's ability to differentiate between daily traffic profiles of days having downed or malfunctioning sensor from days with normal flow values. However, an interesting aspect is that all anomalies clustered according to their GSDs consist of traffic profiles with zero values during high activity periods. Furthermore, zeros occurring during low activity periods are ignored for this group. This shows that the GSDs are sensitive to fluctuations in the traffic profiles.

2.5.6. Switched Subshapes. The results pertaining to the proposed methodology as a means to classify sub-shapes is promising. Specifically, in this section the results of the proposed study that is detailed and explained in Section *Methodoly* are given. Hence, the Tuesdays of 2004 (52 profiles) are separated into two periods am and pm and are illustrated in Figures 2.14a and 2.14b. Figures 2.14c and 2.14d illustrate the joined GSD functions for each set of am and pm periods and an example set of switched GSDs respectively. The ultimate conclusion is that the switched profiles were always clustered into one unique group. In this analysis a variety of clustering methods were used. Specifically, the PAM, Agglomerative and K-Means were used. In addition, each method was tested with a set of distance measures. These measures were Euclidean, Squared Euclidean, Manhattan and Correlation distances. The results were consistent across all the methods, in which all methods under all distances separated the switched GSDs into a unique group, that only contained the switched joined GSD vectors. The methods did differ slightly in their grouping of the original 52 profiles. Simply, the methods were 100% successful in classifying the switched profiles into their own unique group. The interesting aspect of this is that although the switched traces were combined and then clustered highlights that unique GSDs



(a) All Tuesdays of 2004 am period

(b) All Tuesdays of 2004 pm period

(c) Joined GSD functions

(d) Example of switched GSDs

Figure 2.14. The sub shape analysis data

can be created for periods of interest and grouped with days that have similar shapes or behavior. This capability is important in traffic modeling as models could be tailored for specific phenomena or behavior of interest based on period of time or characteristics.

2.6. CONCLUSION

The proposed methodology yields some significant results regarding the classification and clustering of daily traffic flow profiles. The methodology through the employment of Mathematical Morphological operators allows for significant clusters that are stable, tight and statistically significant. Moreover, the proposed methodology is effective in clustering groups with abnormalities into distinct groups. This proves to be extremely useful as data with missing or malfunctioning sensors can be removed prior to the use of the data to parameterize a prediction temporal model as performance of earlier models are found susceptible to outliers. The study amply demonstrate robustness of methodology as inflation in flow profiles does not change the clustering in a significant manner. Its robustness to scaling and shifting also allows for data to be classified not only on shape but also in terms of volume.

In the backdrop of the fact that clustering does not provide perfect answers. The clustering performed herein based on shape shows that besides working days, Sundays and holidays, Saturdays, special days and anomalies have unique shape characteristics and can be grouped distinctively. The results obtained after single step clustering are encouraging in comparison to the past studies where any meaningful interpretation is possible after multistage clustering. The methodology is unique as special emphasis is kept on the clustering design, an aspect missing in earlier studies. The demonstrated capability of GSD function in detecting anomalies alongside classifying separately the days with influence of weather and social events is promising as well as worth exploring. The ability of the SE to decrease computational effort by reducing a single day from 1440 unique observations to 96 without sacrificing the interpretation are clear strengths of the methodology, which can have divergent applications in the traffic domain. The proposed methodology simplifies the pattern analysis as shape covers all features comprehensively. For example, total flows, peak flows, time of the peak flows and off-peak flows are not required to be analyzed separately.

Investigating daily traffic patterns with regard to unique weather shape variations and methodology's potential of detecting missing /erroneous data are likely future extensions. It will lay down a sound foundation for work on a functional prediction model based on groups found through analysis of shape characteristics. In addition, another extension of this work would be to compare prediction models that are parameterized based on shape versus models that use regular clustering. In broader sense, the methodology exploits the interdisciplinary nature of mathematical morphology which is important as clustering of large data sets is an ongoing area of research. For traffic researchers, proposed methodology provides multiple advantages : reduction in computational effort; robustness; detecting sensor malfunctioning and efficient clustering etc all of which are essentially required to solve complex traffic challenges.

3. A HYBRID OF COMPUTATIONAL INTELLIGENCE TECHNIQUES FOR SHAPE ANALYSIS OF TRAFFIC FLOW CURVES

3.1. INTRODUCTION

The ability to accurately analyze traffic flows in an operational setting has been identified as a critical need for Intelligent Transportation Systems (ITS). Previous attempts to accurately analyze traffic flows e.g volume forecasting models etc have been restricted mainly to non-functional approaches and methodologies. Daily traffic profiles display functional characteristics (unimodal and bi-modal curves) and can be more appropriate for functional analysis rather than traditional non-functional approaches. One of the major advantages of functional analysis is that each daily traffic profile is considered as a single datum, which makes it possible to predict on a much longer term or larger horizon with reasonable accuracy. In the traffic classification domain, previous studies used non-functional classification to analyze traffic patterns (Rakha and Van, 1995), (Wild, 1997) and (Chung, 2003). In (Rakha and Van, 1995), two groups were determined which are {Tuesdays, Wednesdays, Thursdays} and {Mondays, Fridays, Saturdays, Sundays}. Furthermore, Wild 1997 classified week days into six groups. The results vary and thus literature lacks consistency regarding classification of days into a clear set or number of groups. Weijermars et al 2007 was the first to suggest the need for studying the shape of the traffic flow profiles. The study concluded that shape of the daily flow profiles may differ between different type of days and motivates the need for such an analysis of traffic patterns shape characteristics. To illustrate this point, consider Figure 3.1, which illustrates two successive Mondays of the same month; however, their shapes differ greatly. In this case Sep. 1, 2004, which is a common holiday (Labor day) in the United States, differs greatly to the other Monday that exhibits the expected bimodal shape characteristics where two peaks are observed corresponding to morning and evening rush hours. To this end, shape based methodology showed that shapes differ within similar days and also has direct interpretation.

The historical literature shows that mainly three approaches are used for traffic analysis: neural networks(NNs); neighbor nonparametric regression and autoregressive integrated moving average (ARIMA) time series models. However, out of these tools neural network is found to be a convenient tool for developing relationships between streams of

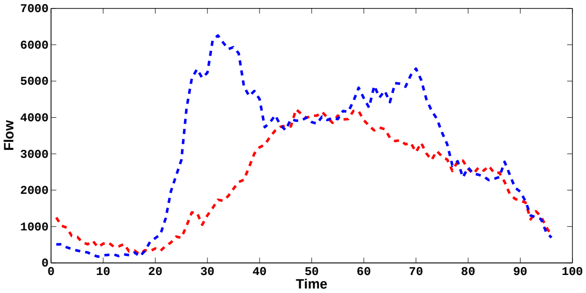


Figure 3.1. Two daily traffic flow profiles for Mondays Sep. 1, 2008 (red) and Sep. 8 (Labor Day), 2008 (blue) are given, which demonstrate unique shapes. Time is in 15 minute interval.

input and output data, not only for pattern recognition to which they are usually associated, but also for a wide range of modeling situations (Kirby et al, 1997). Kirby et al 1997 used a BPNN to carry out a comparative study on NN and statistical models. Neural Networks have performed better than contemporary statistical techniques like discriminant analysis, negative binomial regression, stepwise logistic regression and other classical techniques used in incident detection methodological development (Ivan and Sethi, 1998), (Khan and Ritchie, 1998), gap acceptance modeling (Pant and Balakrishnan, 1994) and safety modeling (Hashemi et al, 1995) , (Chang, 2005), (Sommer et al., 2008). (Hashemi et al, 1995). Comparing NN with discriminant analysis, highlighted that modeling with NNs places no requirements for a specifying a functional relationship and/or indicates their ability to deal with missing data.

As the functional relationship within a neural network is non-linear, it can model undefined, intricate nonlinear surfaces comprehensively, in comparison to many traditional linear statistical models. NNs can effectively analyze the patterns from historical data. Other statistical and mathematical models although proficient in calculation, but are often not effective in predictive analysis as they can not adapt to the irregular varying patterns that can not be written in form of a function. In the field of pattern recognition, NNs classify the patterns from training data and recognizes if the testing data holds the pattern of interest (Patra et al., 2010). In addition, NNs are more responsive to dynamic conditions and do not experience the lag and over-prediction characteristics of time-series models. Owing to the nature of the task at hand, the NN's are considered an appropriate tool for the analysis described herein.

This study conducts a comparative classification analysis of original historical traffic flow data with and without clustering along with GSD transformed data using a back propagation neural network (BPNN) as a baseline. Clustering was done using the Partition Around Medoids (PAM) method with a Gap Statistic as the significance test to optimize the stability of the clusters. PAM is a well established clustering technique which operates on the dissimilarity matrix of input data and minimizes the sum of dissimilarities instead of a sum of squared Euclidean distances. The "Gap" test (Gap statistics) compares the dispersion of clusters generated from the data to that derived from a sample of null hypothesis tests. The null hypothesis test sets are uniformly randomly distributed data in the box defined by the principal components of the input data. GSD transformed data implies the clustering of unique Granulometric Size Distributions generated from each original traffic profile. BPNN is used for training the data's separately and their performance is evaluated by comparing actual and predicted testing targets.

3.2. MATHEMATICAL MORPHOLOGY

Mathematical Morphology is a theory that provides a number of useful tools for image analysis, which is based on the assumption that images consists of structures that can be handled by set theory. In addition, the tools and methods that MM provides can be naturally applied to the analysis of signals. In Gaston-Romeo et al (2011), MM tools were applied to signals in an effort to analyze daily solar radiation time series curves. Similarly, in Guardiola and Mallor (2013) unintended electromagnetic emissions from wireless communication devices are analyzed in the frequency domain. In both previously mentioned works, curves are considered to be bi-dimensional images on which morphological operators are applied to gain information regarding the inherent shape of the two phenomenons. Thus, a time-series curve is considered as a bi-dimensional image. In this study three MM operators are used to study the shape of the daily traffic flow profile. Specifically, Dilation, Erosion and Opening operators. These operators are used to construct a daily traffic profile's *Granulometric Size Distribution* (GSD). The rest of details including mathematical explanation as well as process of obtaining GSDs from original data are detailed in (Guardiola, Wasim and Samaranayke, 2014).

3.3. STUDY DESIGN

3.3.1. Data Details. The study is based on traffic data collected from the I-94 within the Twin Cities Metro area, Minnesota. Data is obtained at station S110 I-94 East Bound/T.H. 65 which has 3 loop detectors D497 94/TH65E1, D498 94/TH65E2 and D499 94/TH65E3. I-94 has 3 lanes in each direction and the station provides composite detector data from three detectors in the eastbound direction. The analysis period is from 1st January 2004 to the 31st of December 2013. However, due to non availability of data due to insensitive detectors, the entire 2009 year data is excluded from the study. Partial data is available for years 2011, 2012 and 2013. The detectors measured and logged the flow for each of the three lanes at 30 seconds intervals. For this study, data is aggregated over 15 minute intervals (96 points per day). Weijermars and Van Berkum 2005 found that 15 minutes data produce better results as the fine grain variations are removed. Unlike past traffic classification studies Rakha and Van (1995), Weijermars and Van Berkum (2005), and Chung (2003) where data consist of 75 days, 118 days and 2 years respectively, this study utilized traffic data of approximately 9 years translating into 2,992 days and 287,232 unique traffic flow value observations at the selected location.

3.3.2. Design of Experiment. The study is designed to ascertain the significance of shape in traffic analysis. The concept is simple, that is, if days are better defined by shape and the shape vary irrespective of the week days than shape based classification and its onward predictive analysis should be better than methods where shape characteristics are not considered. To this end, the initial assumption is made that every day of the week has unique characteristics e.g every Monday of the year is same irrespective of the month and same is true for rest of the weekdays. With this assumption in mind, a simple target of {Mon=1, Tues=2, Wed=3, Thu=4, Fri=5, Sat=6, Sun=7} is created for the entire 2,992 traffic profiles, refer hereafter as the subjective target. The next step is to cluster the original 2,992 traffic flow profiles using PAM with Gap Statistic as the significance test. The resultant 7 clusters become the target representing clustered data and will be referred to as the original target. Lastly, the entire daily traffic flow profiles are converted to their corresponding GSD curves. Clustering is conducted on the GSD curves. The resultant 7 clusters representing unique shapes become the target representing GSD clustered data and hereafter will be referred to as the GSD target. The original 2,992 traffic profiles has three set of targets: subjective target based on initial assumption of every single week day has similar characteristics , original target based on simple clustering results and GSD target obtained by generating the corresponding GSDs of 2,992 traffic profiles followed by clustering. Figure 3.2 explains a simple example with 2004 data used for the supervised BPNN learning (training). The 366 original traffic profiles are selected as input and the subjective target $\{1,2,3,\ldots,7\}$ for 2004 is selected as output. Similarly 2005 data is employed for testing. The trained network is tested for 2005 data and predicted 2005 subjective target values are compared with actual 2005 subjective testing target. The performance is evaluated by percentage of correct classifying target values. Same process is employed for original and GSD cases by training the data's against respective original and GSD targets followed by their subsequent testing. The performance is evaluated similarly by percentage of correct values of predicted target (output) with actual target.

To gain a better insight into the classifying ability of the three cases, a sliding year window methodology is adopted. It implies that the window initially uses one year for training and next year as testing. Subsequently two years are used for training and proceeding year as testing. The process continues until all eight years data is used as training and last year's data (e.g. 2013) is taken as testing. Figure 3.3 depicts the moving year process. The brackets indicate bracketing the years from 1-8, the right side decreasing blocks shows the matching predictions (i.e 1 indicates matching prediction of a year basing on 1 year training data and so on).

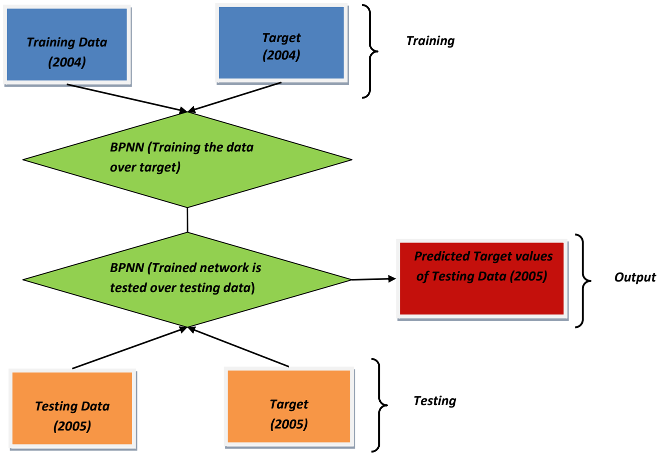

Training Data (2004)
Target (2004)
Training
BPNN (Training the data over target)
BPNN (Trained network is tested over testing data)
Predicted Target values of Testing Data (2005)
Output
Testing Data (2005)
Target (2005)
Testing

Figure 3.2. Diagrammatic layout of BPNN process.

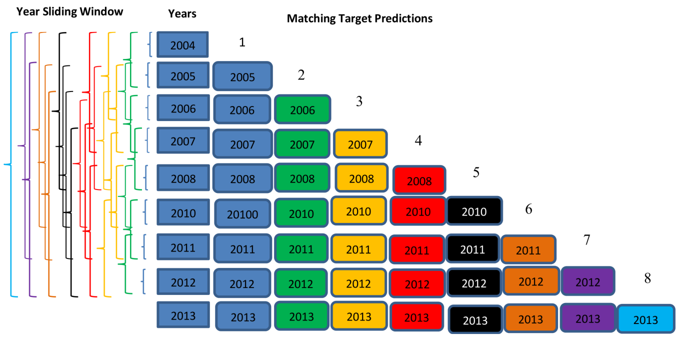
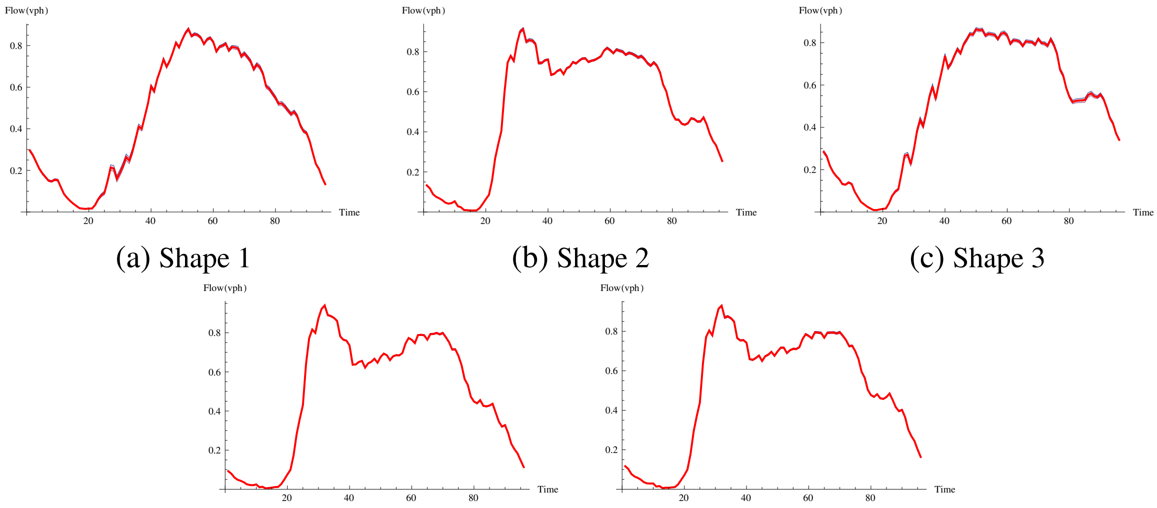


Figure 3.3. Schematic layout of moving year window process.

3.3.3. Clustering Results - Five Basic Shapes . The entire daily traffic flow profiles for 2,992 days are converted to their corresponding GSD curves. Clustering is conducted on the GSD curves using the same method, algorithms and measures. Recall that it is the PAM Algorithm using the gap test statistic method, with a dissimilarity measure of silhouette width to determine the most stable clusters. Out of seven groups obtained, Shape 1 represents Sundays and holidays with typical unimodal shape with peak traffic around 2pm. Shape 3 represents Saturdays, second type of unimodal shape with less significant peak and sustain high traffic around noon to 7 pm. Shapes 4 and 5 represent early and mid week working days behaviors with bimodal shapes. A slight difference exist between the morning and evening peaks of these two shapes. The Shape 2 reflects typical Friday behavior. Although, bimodal in shape yet differs from group 4 and 5 immediately after the morning peak. Figures 3.4a, 3.4b, 3.4c, 3.4d,3.4e represents 95 % confidence interval plotted on five clustered groups. The red thick line shows the mean and therefore define the five distinct shapes obtained through shape analysis of entire dataset of daily traffic flow profiles. The last two groups/shapes representing abnormal behavior (insensitive detector or incidents etc) are not shown as there is no dominant shape.

3.3.4. Back Propagation Neural Network Methodology. Back propagation training algorithm when applied to a feed forward multi-layer neural network is known as Back prorogation neural network. Among back propagation algorithms, Lavenberg Marquardt (LM) is one of the second order methods which overcomes the slow convergence problem and is widely accepted as most efficient in the sense of realization accuracy (Patra et al., 2010). The learning rate is automatically adjusted at each iteration which is termed as the



(a) Shape 1 (b) Shape 2 (c) Shape 3

(d) Shape 4 (e) Shape 5

Figure 3.4. Five distinct shapes representing entire 2992 traffic profiles.

adaptive learning rate. During training, the algorithm takes only 60 percent of the input data for training while 20 percent is used for validation and testing respectively. For every attempt of training, the algorithm selects the data randomly from the whole set and not a fixed set of data. Hence, each time the NN is trained results will differ depending on which 60 percent of the input data is selected for training. As the traffic data is non-linear in nature, sigmoid transfer function is considered more appropriate. Mathematical and further details are not covered as the algorithm is well known among practitioners (refer to (Goh, 1995)).

3.4. RESULTS AND DISCUSSION

3.4.1. Network Architecture. Back propagation neural network is used for comparing predictive classification ability of three approaches. A number of studies analyzing traffic data have used a single hidden layer as it produces satisfactory results. One should refer (Kirby et al, 1997) and (Wei et al, 2007) for examples of applications. Heaton 2008 concluded that problems that require two hidden layers are rarely encountered, but an NN with two hidden layers can represent functions with a multitude of characteristics. However, there is currently no substantial reason to use NN with more than two hidden layers. In the traffic domain, two hidden layers have been used to achieve better results (Ysadi et al., 1999). The literature does not reveal the best approach in determining the number of

Table 3.1. Details of selected NN architecture with single and two hidden layers.

Years	2005	2006	2007	2008	2010	2011	2012	2013	Mean
Single Layer									
Hiddden layer	1	1	1	1	1	1	1	1	1
Hidden neurons	12	12	12	12	12	12	12	12	12
Iterations	20	20	20	20	20	20	20	20	20
Correct class %	77.4	75.9	67.5	67.9	68.4	80.0	66.5	63.3	70.375
Two Layer									
Hiddden layer	2	2	2	2	2	2	2	2	2
Hidden neurons	20-5	20-5	20-5	20-5	20-5	20-5	20-5	20-5	20-5
Iterations	20	20	20	20	20	20	20	20	20
Correct class %	84.2	80.9	77.5	76.4	74.9	81.8	73.7	70.7	76.875

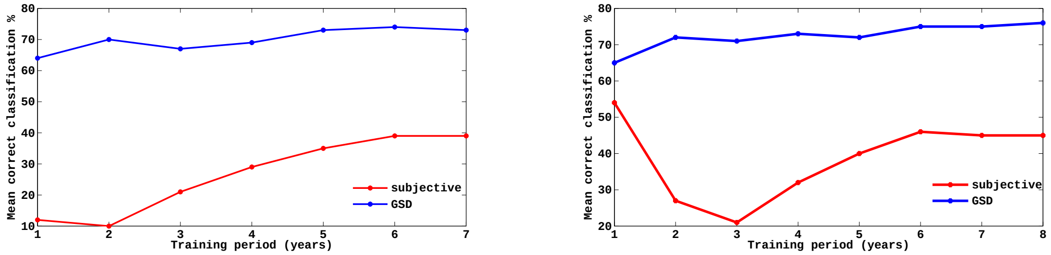
neurons within the hidden layers. However, the number of neurons should be determined in such a way that it results in neither under-fitting or over fitting. A typical value for the number of neurons is 30, which is well supported in the literature as a general rule of thumb (Kirby et al, 1997)and (Wei et al, 2007). Therefore, a range of 5-30 neurons for single as well as two layer architectures is used to search for the optimal number of neurons. The maximum of correct classifying percentage is taken as the criteria to determine the most appropriate network.

A MATLAB code employing the *newff* function is executed to select the optimal number of neurons using same training data-set. For the single layer architecture, the best performance was found to be one with 12 neurons in the hidden layer with an average correct classifying percentage of 70.375%. In order to get the best two layer architecture, all possible combinations of neurons in two hidden layers ranging from $5-30$ neurons are tested. The best performance is found with the combination of 20 and 5 neurons in two layers. The mean correct classification percentage improved significantly to 76.875%. Therefore, the selected Multilayer Perceptron (MLP) architecture is two hidden layers with 20 and 5 neurons respectively. Table 3.1 demonstrates the relative performance of the two architectures (e.g. single and double hidden layers MLPs architecture).

3.4.2. Comparing Subjective and GSD Output. Firstly, the question of whether every day of the week has a unique shape or days differ in shape owing to the functional and behavioral characteristics is addressed. If shape of every day of the week is unique and remain as such, then subjective target should have better classification performance. If the assumption made above is not correct and shape does change, then the GSD output which is solely based on shape of the traffic profiles should perform much better than the original output.

Table 3.2. Comparison of subjective and GSD based on mean correct classifying percentage

Years/	2005	2006	2007	2008	2010	2011	2012	2013
Training period	%	%	%	%	%	%	%	%
1 year	61(76)	62(78)	56(66)	58(70)	9(66)	61(72)	12(64)	54(65)
2 year		66(80)	61(70)	61(75)	10(68)	34(74)	10(70)	27(72)
3 year			65(73)	65(78)	10(72)	22(79)	21(67)	21(71)
4 year				64(76)	10(72)	18(78)	29(69)	32(73)
5 year					9(73)	18(77)	35(73)	40(72)
6 year						18(78)	39(74)	46(75)
7 year							39(73)	45(75)
8 year								45(76)



(a) Classification performance - 2012.

(b) Classification performance - 2013.

Figure 3.5. Classification performance of original and GSD.

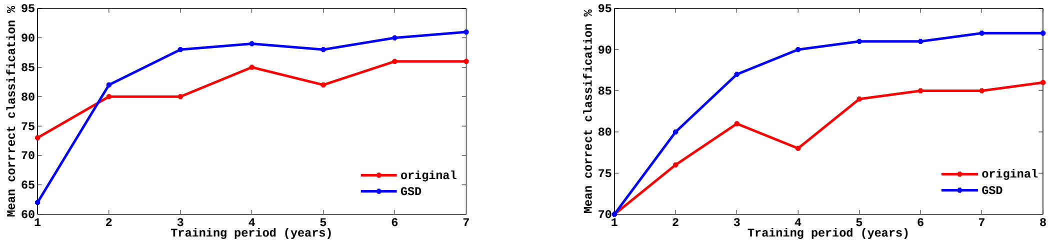
A comparison is performed between subjective and GSD output for whole 9 years data with moving year sliding window and 30 training iterations. The input consists of 96 rows and 2,992 columns matrix of traffic data while the output consists of 96 rows of 1 column matrix. The training data is trained and tested initially with the subjective target and later with the GSD target. The result indicates that GSD classification performance is superior to the subjective output. Table 3.2 demonstrates the relative classification performance of subjective and GSD outputs (in brackets) respectively. Figures 3.5a and 3.5b illustrate the relative classification performance for 2012 and 2013 respectively. It is interesting that GSD maintain a steady classification performance throughout all the years and improved with the increase in the amount of training data. Furthermore, a maximum correct classification of 73% with 8 years training input is achieved. It implies that with more training data, the NN results are improved to a certain level as it gets more training experience with shape profiles and recognizes better. It demonstrates that shape is not a static phenomenon rather its dynamic and varies within days of the week.

Table 3.3. Comparison of original clustered and GSD clustered targets, based on mean correct classifying percentage.

Years/	2005	2006	2007	2008	2010	2011	2012	2013
Training period	%	%	%	%	%	%	%	%
1 year	80(82)	78(76)	72(70)	77(74)	73(73)	80(84)	73(68)	70(70)
2 year		85(85)	77(78)	79(85)	76(80)	80(89)	80(82)	76(80)
3 year			80(85)	82(85)	81(84)	84(91)	80(88)	81(87)
4 year				84(87)	81(86)	87(91)	85(89)	78(90)
5 year					83(89)	88(93)	82(88)	84(91)
6 year						88(93)	86(90)	85(91)
7 year							86(91)	85(92)
8 year								86(92)

3.4.3. Comparison of Original and GSD Output. In the preceding section 3.4.2, the significance of shape is further validated. In this section, the output of original target trained on original data with the corresponding GSD output trained on GSDs of same original training data are compared. The comparative analysis is of practical form, as clustering is a common procedure employed when developing traffic prediction or incident detection models. If GSD is better in classification performance than the original output, then it should demonstrate that shape can contribute significantly in improving traffic analysis results and thus suggests that functional approaches are a better option than the prevailing non-functional approaches.

The input consists of 96 rows and 2,992 columns matrix of data for original target and same size matrix of corresponding GSDs for training with GSD target. The training data is trained and tested initially with the original target and later with the GSD target. The results indicate that GSD classification prediction is superior to the original. It shows that days differ in shapes and functional characteristics can not be ignored in traffic analysis. Table 3.3 and Figures 3.5a and 3.5b demonstrate and illustrate the relative classification prediction performance of original and GSD outputs(in brackets) respectively. In case of 2012, initially the performance of original target is better than GSD, however with the increase in the amount of available training data the GSD has better classification prediction performance. In the second case, initially both performances are equal but with increase in training data the GSD again gets better results. Another aspect is the initial rise in GSD performance which becomes stable after being trained for 2 years in 2012 and for 4 years in 2013. It demonstrates that $2-4$ years training data is sufficient to train the NN on traffic flow profiles and any further data will not significantly improve the BPNN results.



(a) Classification performance - 2012.

(b) Classification performance - 2013.

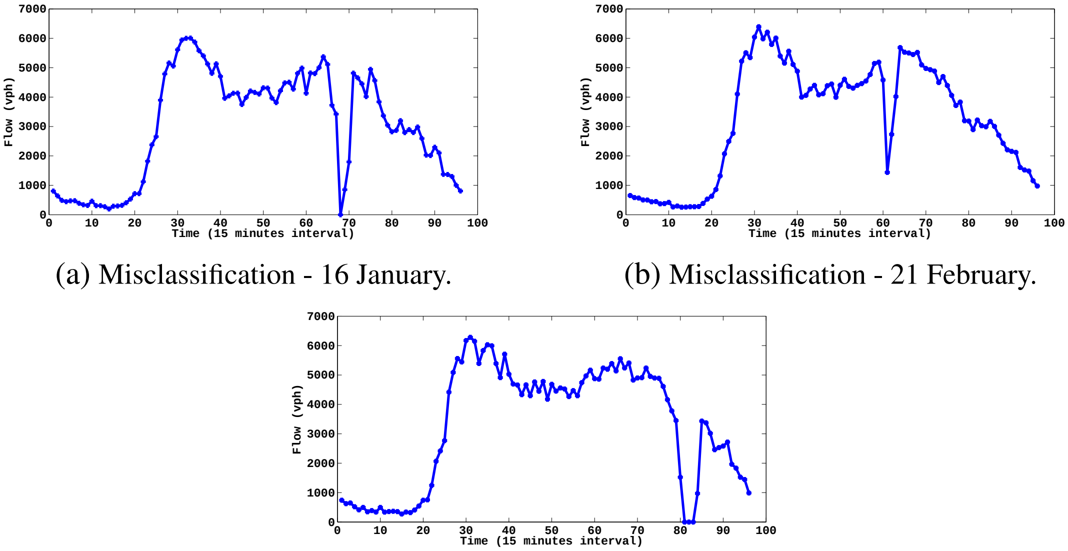
Figure 3.6. Classification performance of original clustered and GSD clustered.

3.4.4. Analyzing Misclassifications. A confusion matrix is a visual performance assessment of a classification algorithm. To this end, confusion matrices are computed to analyze miss-classified days by the BPNN based on results obtained in section 3.4.3. Classification prediction of 2013 is carried out based on 8 years worth of training data from (2004-2012) after satisfactory BPNN training. The best outputs of ten training iterations is selected for the original and GSD ensuring that the mean correct classification percentage is in close proximity of mean values already obtained (refer Table 3.2). The resultant two confusion matrices with mean correct classification percentage of 85.43% for original groups represented by ‘G’ and 92.23% for GSD shapes represented by ‘S’ are depicted in Table 3.4. The total number of misclassifications observed are 24 for GSD and 45 for original. The balanced confusion matrices also suggest that the BPNN architecture is satisfactory. This finding that shape is important is valuable in the analysis of traffic.

Cases of misclassifications are discussed to explain the relative performance of shape analysis over traditional besides explaining the glaring instances of BPNN failure. Figure 3.7a represent traffic profile of 16 January 2013, which was a working day and defined with conventional morning and evening peaks. However as evident from the figure it has a single insensitive detector reading at interval 68. Owing to this abnormal shape GSD has classified it along abnormal/insensitive detector shapes while the traditional clustering placed it along non working days (Sundays). BPNN classified it along working days not taking into account a major shape deformation. Similarly, the traffic profile of 21 February depicts a working day profile with the exception of dip which might be due to incident or an insensitive detector, The GSD method classified it along abnormal behavior/insensitive profiles, while traditional classified it as non working days (Sunday) and BPNN misclas-

Table 3.4. Confusion matrices - original clustered and GSD.

Groups	G1	G2	G3	G4	G5	G6	G7		Shapes	S1	S2	S3	S4	S5	S6	S7
G1	9	3	0	0	0	0	0		S1	73	1	0	1	3	1	0
G2	0	0	0	0	0	2	0		S2	0	36	1	0	3	1	0
G3	6	1	64	0	0	7	4		S3	0	0	11	0	0	0	0
G4	0	1	0	50	0	2	0		S4	0	0	0	51	0	1	1
G5	0	0	0	1	45	0	0		S5	5	0	0	1	69	1	1
G6	1	3	5	0	0	75	1		S6	0	0	0	1	0	42	0
G7	3	0	1	0	0	4	21		S7	0	0	2	0	0	0	3



(a) Misclassification - 16 January.

(b) Misclassification - 21 February.

(c) Misclassification - 7 March.

Figure 3.7. Misclassification examples - 2013

sified it as working day. Figure 3.7c illustrates the traffic profile for the 7th of March (a working day), however, the shape does not reflect common working day behavior rather represents a non-working day shape along with a dip near the end of the day. Due to shape exhibiting non working day behavior, GSD classified it as non-working day while BPNN predicts it as working day.

The analysis of misclassified cases shows that shape of traffic profiles really change and not constant for every day. The shape based GSD is quick in identifying an abnormality in shape where traditional methods fail. It also depicts that BPNN although performed reasonably well in evaluating normal shapes; however, its performance is questionable in cases where shape exhibits deviated behavior from standard shapes (abnormal shapes due to incident and insensitive detectors).

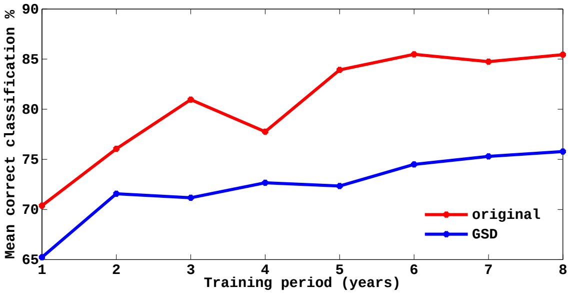
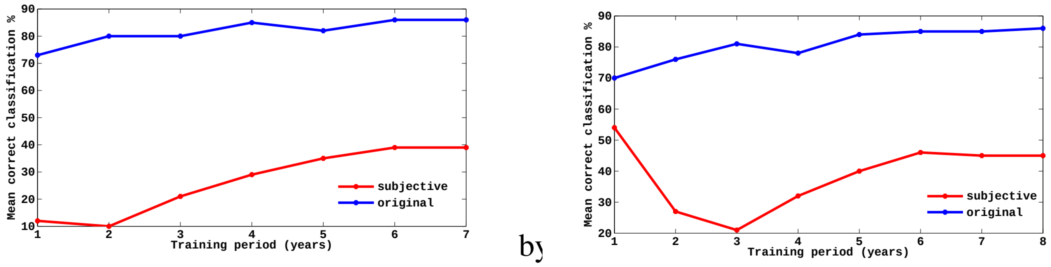


Figure 3.8. Comparative performance of original and GSD for same original training data.

3.4.5. Performance of Original and GSD Targets Trained on Original Traffic Profiles. Another comparison is performed by training GSD target on original data rather on GSDs and comparing it with results of original target trained on same data already obtained in (refer section 3.4.3). The classification prediction performance of clustered is found better in almost all types of yearly training. The performance gap is narrow initially but becomes wider with an increase in training data with the exception of 4 years worth of training data (refer to Figure 3.8). One understandable reason for GSD having a lower performance is that GSD targets are obtained by clustering GSD profiles and not original profiles. It shows that if shape based MM methodology is to be employed then traffic data has to be dealt in terms of GSDs and not original traffic flow time series curves.

3.4.6. Validating Clustering Prior to Prediction. Clustering explains the hidden structure within the data and provides a simple but meaningful description of data distribution. Irrespective of the prediction algorithm used, the prediction accuracy would be affected if data is not fully understood and processed. This study validates clustering is a necessity as well as its quality prior to prediction. In section 3.4.2 data with subjective target of arbitrary values $\{1,2,.....,7\}$ is analyzed. Although no actual clustering is carried out but yet some partition within the data basing on arbitrary values is considered. The predicted output of data with subjective target (arbitrary target values) is far below the predicted output of original target refer figures 3.9a , and 3.9b. It is observed that partition of the data and minimizing the distances between the data points with respect to the center point obtained through clustering helps in better classification and prediction. Apart from



by

(a) Performance - 2012.

(b) Performance - 2013.

Figure 3.9. Classification prediction performance of clustering vs non clustering

minimizing the distances within data-set, clustering also contributes towards complexity reduction in the NN due to high similarity all the data and this contribute in enhancing predictive accuracy.

3.5. CONCLUSION

This study concludes that shaped based analysis, clustering and prediction are all substantially increased in performance through the employment of shaped based classification. The study highlights the significance of functional over presently used non functional approaches in analyzing traffic data. The comparison between original data (clustered and non clustered) and GSD transformed traffic profiles demonstrate efficacy of shape in classification and prediction. The results show that MM provides a more stable shape based clustering that classifies the existing shape patterns from the training data and recognizes it efficiently during testing.

A major contribution is that shaped based classification has the potential of the methodology to improve the existing traffic prediction models performance by employing shape based clustering prior to the development of a prediction model. Apart from emphasizing the significance of shape, the study also highlighted the necessity of clustering prior to traffic analysis. The performance of BPNN remain satisfactory especially in the context of data used in this study. It is found that $2-4$ years of training data is sufficient for training and any further addition does not improve results significantly. MM tools such as the GSD is one of the techniques in practice for shape analysis, however, other contemporary techniques are also required to be explored. Investigating the functional model to predict traffic profiles is a likely future extension of this study.

4. A FUNCTIONAL TIME SERIES APPROACH FOR TRAFFIC FORECASTING

4.1. INTRODUCTION

Traffic flow forecasting has been regarded as a major concern for many of the current Intelligent Transportation Systems (ITS). In addition, efficient traffic flow forecasting assists in the development of management strategies in Advance Traffic Management Systems (ATMS) and Advance Travelers Information Systems (ATISs). Furthermore, it supports future decision making regarding expansion of traffic facilities, modification to roadway networks, and other key decisions related to increasing the quality of service to its users. In Zheng et al. (2006), it is argued that the ability to make continuous predictions of traffic flows and link travel times for even a few minutes into the future, using real-time traffic data, is the primary requirement for providing dynamic traffic control and guidance. As noted in Dia (2001), predictive information can be divided into two distinct categories: strategic and short-term. Short-term traffic flow prediction is useful to effectively deal with the dynamic traffic situations arising from congestion, accidents, and weather related problems. The strategic or long-term flow prediction contribute only marginally to the control and management arising from fluid traffic problems; however, it contributes in the planning of potential growth in traffic volume, which arises from urban sprawl or changes in infrastructure.

The technological developments in traffic data collection and storage techniques in recent decades, have allowed more dense sampling of observations over time and space. These data collection and storage techniques allow researchers to observe and record real-life processes in great detail, unlike any time in the past Kidzinski (2015). Examples of this growth in data expands across all traditional disciplinary boundaries and fields of study. Specifically, the propagation of very large data sets due to the availability of sophisticated data gathering and storage techniques has led to term *"big data,"* which can be found in financial transaction data, satellite photography, pollution levels and distribution in time and so on. In traffic, this large data can be found along with the proliferation of censoring and monitoring of traffic systems, which collect video, count data, and weather conditions to name a few. Due to the amounts of data and its high dimensionality, classical statistical tools become inadequate and inefficient Donoho et al. (2000). One of the prominent new technique that overcomes this inadequacy is Functional Data Analysis (FDA).

In analyzing traffic data, the functional approach can be interpreted as reflecting the influence of certain smooth functions that are assumed to underlie and to generate the observations. The traditional multivariate statistical approach ignores vital information about the smooth functional behavior of the generating process that defines the data Green and Silverman (1994). In particular, functional data analysis can often extract additional information contained in the function and its derivatives that is not normally available to traditional statistical methods. The functional approach treats the whole traffic flow profile over a given day (curve) as a single datum. The basic idea behind functional data analysis is to express discrete time series observations in the form of a function that represents the entire measured function as a single observation or datum. This represents a change in philosophy towards the handling of traffic time series data and provides a motivation to consider functional time series approach for traffic forecasting. In Kargin, Vladislav and Onatski (2008), a thorough explanation of the underlying theoretical concepts of functional data analysis undertaken herein. Furthermore, Kargin, Vladislav and Onatski (2008) proposes the use of predictive factor technique that aims to forecast curves through the employment of functional autoregression.

Traffic flow forecasting has been extensively investigated for more than two decades. Numerous techniques have been employed in the context of traffic flow forecasting, depending upon type of data and the potential end use of the forecast. These techniques include but not limited to: time series models William et al. (1998), Lee and Fambro (1999), Williams (2003) and Stathopoulos and Karlaftis (2003), neural network based models Chen and Grant (2001), local linear regression methods Sun at al (2003), Kalman filtering techniques Zheng et al. (2006), Xie, Zhang and Ye (2007) and Cetiner et al. (2010) and fuzzy neural models and fuzzy logic system methods Yin, Wong, Xu, and Wong (2002) and Zhang and Ye (2008), parametric regime switching space-time model Kamarianakis (2012), visco-elastic models Zhu and Chun (2013), and multiple kernel learning and vector support machine Yu and Lam (2014). Recently, Chiou et al. (2012) adopted functional approach and proposed a stochastic functional mixture model for predicting traffic flow. In Guardiola, Wasim and Samaranayke (2014), FDA is used to analyze patterns of daily traffic flow curves as a means to develop traffic flow monitoring and control mechanisms; however, no study of prediction is developed.

In the literature, no clear durations are defined for traffic prediction horizons and generally the prediction horizon for short-term is limited to 30 minutes while long-term predicts days, weeks, and months ahead. In between the short and long-term horizons, medium-term prediction horizon is introduced in this study, which is defined as forecast horizon of more than 30 minutes and for this study one hour ahead is considered. Currently, there is no feature to obtain driving directions based on traffic considerations at a specific future date and time embedded within the most popular GPS navigation systems and devices such as Google Maps, TomTom, or Smart phone software applications. These devices and application supply driving directions mostly with algorithms that seek to minimize travel distance and supply notification of current driving conditions but can not supply directions in future time with conditions in the future incorporated in the way they optimize a user's route.

The contribution of this research presented herein is the use of functional time series approach in forecasting traffic flow in short and medium-term horizons using real-world data from loop detectors. The functional time series formulation we employ is introduced in Shang (2013) and is a simpler variation of the more general functional time series discussed in Horvath and Kokoszka (2012). This functional concept is simple but innovative as in-spite of using data points individually, each traffic profile over a given day is considered as a single datum. Future functions are forecasted based on the principal component scores. The one-step ahead forecast produced by this method yields a complete next day forecast, which provides a predicted traffic profile for the entire day. Similarly, incase of partially observed traffic profiles, the functional context implies forecast for the remainder day. However, in this study, only one hour ahead or four forecast points from the complete day or remainder day are considered. The remainder of this study is organized as follows. In Section *Functional Time Series Model*, the underlying model is explained. A brief description of the data used to validate the proposed application of functional time series is given in Section *Data Details*. Lastly, the results/discussion and conclusion are given in Sections *Results* and *Conclusions* respectively.

4.2. FUNCTIONAL TIME SERIES MODEL

Functional time series treats observations as realizations of a function observed at discrete points in time. Models that incorporate this structure of a continuous functional form from which the data is generated can lead to more precise and meaningful findings as the error brought about from interpolation is reduced. Functional time series analysis has begun to appear as an efficient means of analyzing time series as seen in works such as Hovath, Huskova and Kokoszka (2010), Hormann and Kokoszka (2010), and Hyndman and Shang (2010) but this trend has begun very recently in-spite of the fact that Aguilera et al. (1999) proposed functional principal components regression (FPCR) to model and forecast functional time series two decades ago. The concept of FPCA is successfully demonstrated in different fields of applications, such as electricity demand forecasting Antoch J. (2008), breast cancer mortality rate modeling and forecasting Erbas et al. (2007), call volume forecasting Shen and Huang (2008), climate forecasting Shang and Hyndman (2011), and demographical modeling and forecasting Hyndman and Shang (2009). However, the approach has remained relatively unexplored within the traffic domain.

The functional time series formulation employed in this research is based on the methodology explicitly described in Shang (2013) and is dependent on the use of Functional Principal Components (FPC). There are more general ways of modeling functional time series Kargin, Vladislav and Onatski (2008), but the approach employed in this study is relatively easy to implement in the sense that the time dependence between successive functions is modeled using traditional univariate time series methods. A major aim of this research is to develop a methodology for predicting the $24 - hour$ traffic flow curve (profile) of a future day, based on the flow curves observed on n successive past days. It is assumed that there is a smooth function f_t that describe this profile for a given day for $t = \{1, 2, \ldots, n\}$. The set of functions $\{f_1, f_2, ..., f_n\}$ forms the observed portion of a functional time series. Following a formulation very similar to that in Shang (2013), it is assumed that the functional values $f_t(x)$, $t = \{1, 2, ..., n\}$ are not directly observed but instead the quantities $z_t(x_i)$ are observed, where $z_t(x_i)$ equals the total flow over the short time interval $(x - \Delta x)$, x denotes a time point on a day t and Δx signifies the span of the interval over which the traffic flow is measured. In this study the value of Δx was set to 15 minutes and the traffic flow profiles were observed over 84 days, but only data from the first 83 days were used for estimating the predictive model $(n = 83)$. The $z_t(x_i)$ values are employed to obtain smoothed functions f_t using the relationship

$$z_t(x_i) = f_t(x_i) + \sigma_t(x_i)\varepsilon_{t,i} \tag{2}$$

where, $\varepsilon_{t,i}$ are independent identically distributed (*iid*) random variables with zero mean and unit variance, and the x_i denote time points 15 minutes apart given by $x_i = 15i$, for $i = \{1, 2, ..., 96\}$. Note that the $\sigma_t(x_i)$ are multiplicative factors that enable the noise components $\sigma_t(x_i)\varepsilon_{t,i}$ to have different variances across t and i. The goal is to forecast the values $z_{n+h}(x_i)$, over $i = \{1, 2, ..., 96\}$ for a future *24-hour* period h days ahead of the last observed day n.

Given a realization f of a random function defined over a compact domain $[0, \tau]$, we can write

$$f = \mu + \Sigma_{k=1}^{\infty}\lambda_k\upsilon_k \tag{3}$$

where, μ is the population mean function, λ_k is the k^{th} principal component score, and υ_k is the k^{th} population functional principal component Shang (2013). The functional principal components are the normalized eigenfunctions of the population covariance operator associated with the underlying random function Horvath and Kokoszka (2012). In this context, the smoothed daily traffic flow curves f_t obtained from the relationship expressed in Equation 2 can be written as,

$$f_t(x) = \bar{f}(x) + \sum_{k=1}^{K} \hat{\lambda}_{t,k}\hat{\upsilon}_k(x) + \eta_t(x), \tag{4}$$

where, $\bar{f}(x) = n^{-1}[\sum_{t=1}^{n} f_t(x)]$ is the sample mean of the functions $f_1, f_2, ..., f_n$ evaluated at the time point x, $\hat{\upsilon}_k$ is the k^{th} orthonormal eigenfunction of the empirical covariance operator and $\hat{\lambda}_{t,k}$ is the k^{th} principal component score associated with the traffic flow function of that day t. The value of K is chosen using a scree plot and represents the optimal number of principal components that can adequately describe the behavior of the underlying functions. The term η_t represents the error in approximating $f_t(x_i)$ by $\bar{f}_x + \sum_{k=1}^{K} \hat{\lambda}_{t,k}\hat{\upsilon}_k(x)$

Let $\daleth(x)$ denote the set $\{f_1(x), f_2(x), ..., f_n(x)\}$ of smooth functions extracted using Equation 2 and let $\bigwedge$ denote the set $\{\hat{\upsilon}_1, \hat{\upsilon}_2, ..., \hat{\upsilon}_n\}$ of eigenfunctions estimated from the empirical covariance operator. For each k, the principal component scores $\hat{\lambda}_{t,k}, t =$

$1, 2, ..., n$ can be treated as a univariate series. Standard time series techniques can be employed to obtain predicted future values $\hat{\lambda}_{n+h,k|n}$. It is worth noting that vector time series methods are not needed because $\hat{\lambda}_{t,k}$ is uncorrelated to $\hat{\lambda}_{t,s}$ for all $k \neq s$, for all values of t. Following Shang (2013) it can be shown that the h step ahead predictor of $z_{n+h}(x)$ based on $\daleth(x)$ and $\bigwedge$ is given by the conditional expectation

$$\hat{z}_{n+h}|_n(x) = E[z_{n+h}(x)|\daleth(x), \bigwedge] = \bar{f}(x) + \sum_{k=1}^{K} \hat{\lambda}_{n+h,k}|_n \hat{\upsilon}_k(x), \tag{5}$$

where, $\hat{\lambda}_{n+h,k}$ are the predicted values from the univariate time series of $\lambda_{t,k}$ time series.

4.3. STUDY DESIGN

4.3.1. Data Details. The data used in this research consists of traffic flow profiles collected from the I-94 passing through the Twin Cities Metro area, Minnesota. Data is collected at station S110 I-94 East Bound/T.H. 65, which has 3 loop detectors named D497 94/TH65E1, D498 94/TH65E2 and D499 94/TH65E3. The I-94 is an urban freeway, which has 3 lanes in each direction and the station provides composite detector data from three detectors in the eastbound direction (Refer to Figure 4.1). The analysis period is from the 1st of January to the 24th of March 2004, thus totaling 12 weeks of data. The twelve weeks data is comprised of 84 days in total, out of which 83 days are used for analysis while the 84th day is utilized for forecasting. The detectors measured and logged the flow for each of the three lanes at 30 seconds intervals. For this study, data is aggregated to 15 minutes intervals (96 points per day). Previous research has found that 15 minutes data produce better results as the fine grain variations are removed refer to Weijermars and Van Berkum (2005).

4.3.2. Methodology. To highlight the functional time series approach in the traffic domain, two types of functional approaches are employed. First, a simple functional time series model based on Functional Principle Component decomposition is developed to forecast a complete day's traffic profile. In the second method, the same model is employed; however, a Penalized Least Square (PLS) method is used to dynamically update a partially observed traffic profile. To compare the performance of the functional time series approaches, the results obtained are directly compared to the commonly employed approach of using traditional autoregressive moving average (ARMA) for forecasting traffic flow. To ascertain the relative performance of the methods for different flow conditions, three conditions are selected representing the flow observed during late night, morning peak, and evening peak periods. Although a complete day ahead forecast is obtained through the functional methods, for comparative analysis purposes, only four steps ahead forecast horizon is considered (a step corresponding to 15-minutes). This translates into a total of one hour ahead forecast. For the purposes of clarity, it should be noted that all functional methods provide a full daily profile ahead; however, traditional ARMA methods can only do a few points ahead. Thus, a comparison of the ARMA results to the Functional methods employed would be deemed unfair if the horizon is too far into the future. This is due to the prediction intervals increasing as the prediction horizon is increased for ARMA methods. It is well known in previous stated literature herein that ARMA methods perform well in a horizon of 1 hour. Hence, it is the goal of the study to compare ARMA methods in a horizon where they have been shown to outperform other methodologies (these works are highlighted in the latter ”Forecasting through ARIMA” section) . A relative comparison is based on four performance measures namely the root mean square error (RMSE), mean absolute error (MAE), mean absolute deviation (MAD), and mean absolute prediction error (MAPE) respectively.

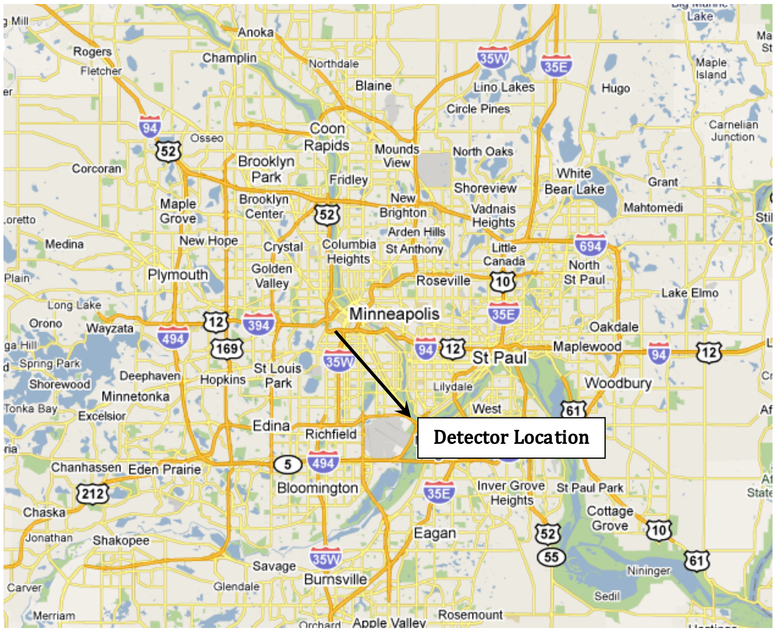


Figure 4.1. Map of Minneapolis MN U.S.A Highway and Freeway system. Arrow indicates position of detectors on I-94.

4.4. RESULTS

4.4.1. Forecasting Through Functional Time Series Model.

When all traffic profiles are complete, FPCA allows a decomposition of the data into a number of functional principal components and their uncorrelated scores. The functional principal components explains where the largest variation lies within the data while the corresponding coefficients of functional principal components, which are termed as scores, describe the magnitude of the variation. The main idea is to take into account the continuous feature of the data. The task of converting the functional data into corresponding smooth functions can be accomplished either through interpolation or smoothing of the data. Data is collected at 30 seconds intervals and then aggregated at a higher interval of 15 minutes, which automatically removes the fine grained variation within the data. This aggregation therefore produces a sufficiently smooth curve on which no additional smoothing procedures are needed. The first step is therefore to find the amount of variation explained by the functional principal components (FPCs) apart from the mean function. A *scree plot* is commonly used to determine the appropriate number of FPCs to be retained in in the model. Figure 4.2 shows that first three FPCs explain a substantial portion of variation within data. Table 4.1 shows that the first three FPCs aggregate to 93.5% of the total variation and hence are retained for the model. It is common to keep the number of FPCs that explain at least 90% of the variation. Thus, retaining only the first three FPCs for further model building satisfies this condition.

Table 4.1. Variance proportion explained by FPCs

FPCs	1	2	3	4	5
Variance proportion	79.8615%	11.755%	1.8623%	1.164%	0.67

Figure 4.3 and 4.4 illustrate the selected three functional principal components and their associated PCA scores. The top left graph in the Figure 4.3 illustrates the mean function, while the remaining figures illustrate the three functional principal components. The first FPC which explains 79.86% of variation represents two peak flows or rush hours with high-volume periods from 7-10 AM and 3-7 PM. An analysis of first FPC shows

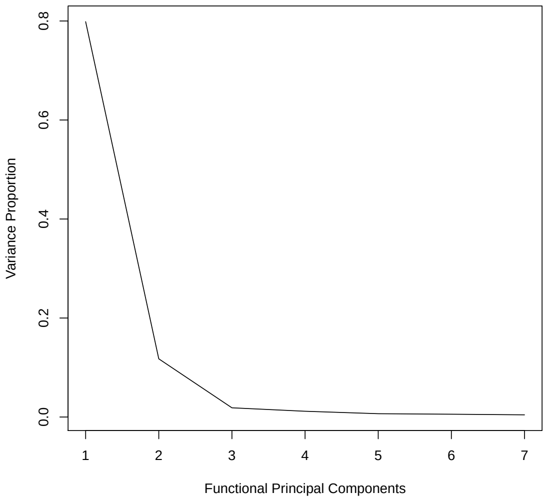


Figure 4.2. Scree plot.

that the high peak in the morning suggests that maximum variation is observed within the morning peak or in other words the flows in the morning peak fluctuates more than the other parts of the day. The second FPC which explains 11.75% of variation appears to be related with low flow periods after the PM peak and traffic flow between morning and evening peaks. This variation is related to typical traveller's behavior on non working days due to a shift in the conventional morning peak from 7-9 AM to 11 AM-3 PM. Similarly, the traffic flow after PM peak also differs from normal working days. The third FPC explains only 1.8623% of the total variation and represents the low volume abnormal behavior during the day time besides some high-volumes observed at the late night period. Figure 4.4 plots the forecasted principal components scores. The associated scores defines the magnitude of variation. The interpretation of the three FPCs above is reinforced by their corresponding FPCs scores as in first case the cyclic behavior reflects that magnitude of variation is more in case on working days as higher scores are related to working days. In the second case, the higher scores are related to the non-working or weekend days and is represented by peaks occurring after successive intervals. The peaks represent high score on weekends with in between low scores assigned to the normal working days. While in last case the higher scores are related to days with abnormal behavior of low flows. Figures 4.5a and 4.5b illustrate the one step ahead prediction and its 80% confidence intervals. In the functional context, one step ahead forecast implies predicting the complete next day's traffic profile. In Figure 4.5a, the red traffic profile predicts the the next day's (i.e. 24th of March (a Wednesday)) traffic profile while the black traffic profiles in the background depict the data from which that prediction was obtained by training the model.

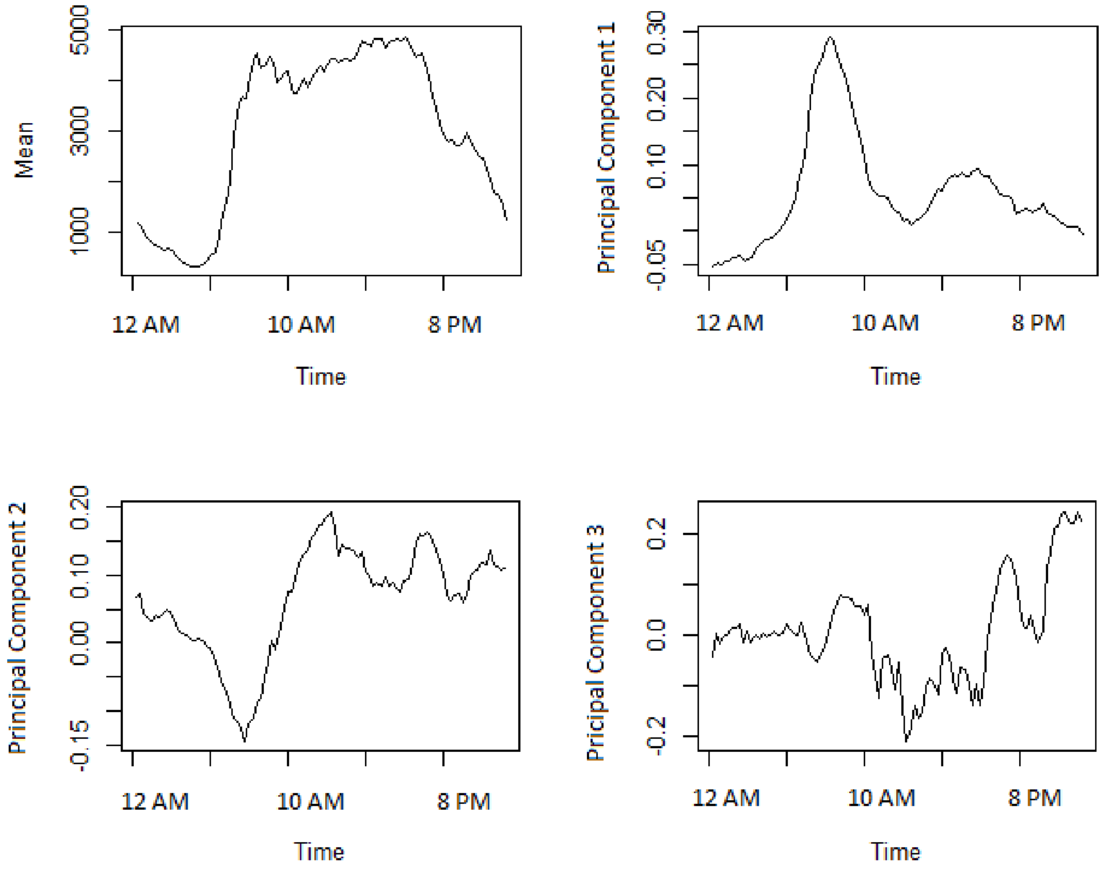


Figure 4.3. First three Functional Principal Components (FPCs).

In traffic domain it is equally important to forecast the partially observed traffic profile by updating the point forecast. In fact, every daily segment of a traffic profile can be treated as a univariate time series. To improve the forecast accuracy, it is desirable to dynamically update the point forecast for the rest of the most recently observed traffic profile, in other

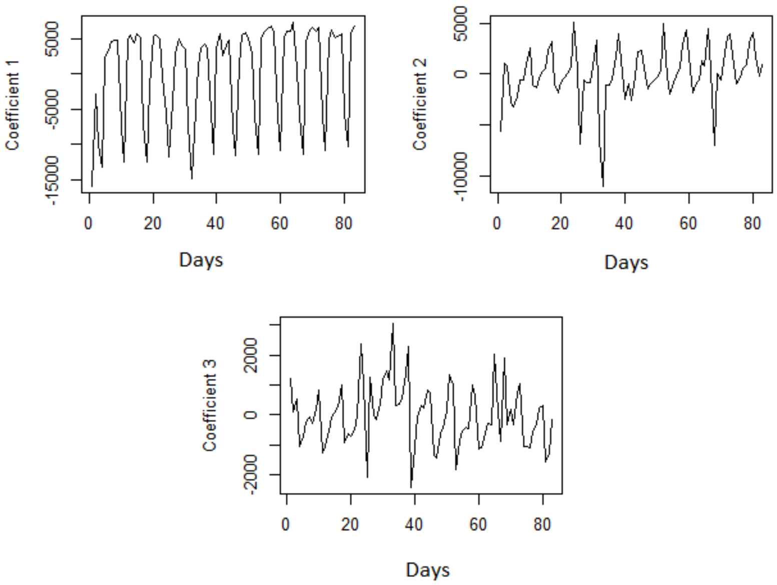
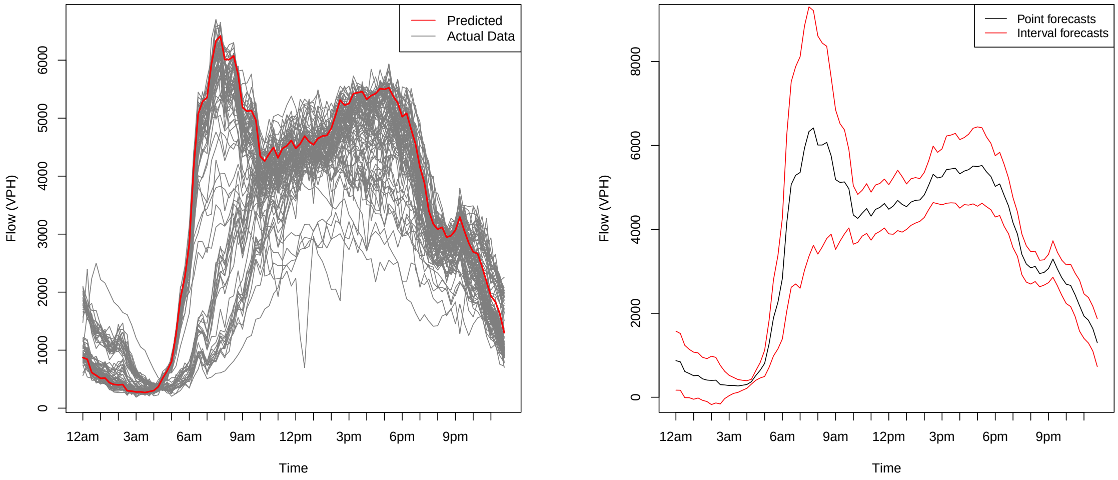


Figure 4.4. Associated principal component scores of the three FPCs.

words, near real-time traffic prediction. This is significant in the cases where a partially observed traffic profile has influence on the remaining part of that profile (e.g. the first half of the traffic profile has influence on the remaining day's profile or the AM peak has an impact on the PM peak). For dynamic point and interval updating, a regression based on Penalized Least Square (PLS) is used for dynamic updating, when partial data in the most recent curve is observed and the requirement is to forecast some intervals ahead or the remaining part of the observed daily traffic profile. The technique is adequately explained by Shang (2013) and details are omitted. In the PLS method the updating is carried out by using a regression based approach in which regression coefficients are estimated by minimizing a penalized sum of squares. Penalized least squares estimates provide a way to balance fitting the data closely and avoiding excessive roughness. The results obtained by applying this technique on our data with last day 24th of March, which was sliced into equally divided halves with one representing the traffic profile from midnight to noon (AM part) and other from noon to midnight (PM peak) and are illustrated in Figure 4.6. The AM peak remained while the PM peak is removed from the data and is forecasted



(a) Forecast for March. 24th (Wed), 2004.

(b) Forecast with 80% confidence intervals.

Figure 4.5. Functional one step ahead forecast (24th March, Wed) with 80% confidence intervals.

using the PLS method. An analysis of Figure 4.6 shows that the PLS forecast provides a satisfactory approximation of the true partial traffic profile. That is, the PM period is predicted and closely resembles the original PM period that was removed. It can be seen that the confidence intervals are wide from 3–6 PM, while confidence intervals narrowed after 6 PM. This reflects the model performance in forecasting traffic between 3–6 PM is less reliable due to more variability in the observed traffic as compared to forecasting traffic past 6 PM, which is quite predictable. Hence, the forecasting methodology produces a curve that closely resembles the original data.

4.4.2. Forecasting Through ARIMA. From the traditional time series perspective, ARIMA is the most widely used time series analysis method, which aims to determine the regression type relationship between the historical data and the future data. In AR models the regression type relationship is exploited. In ARIMA, its more complicated as we exploit the linear relationship which is an infinite AR. That is to say that ARMA models assume constant variance, where ARIMA models have infinite variance. ARIMA as well as its derivatives has been widely applied to model many types of time series, including traffic flow series and they have become indispensable tools for short-time prediction Ahmed

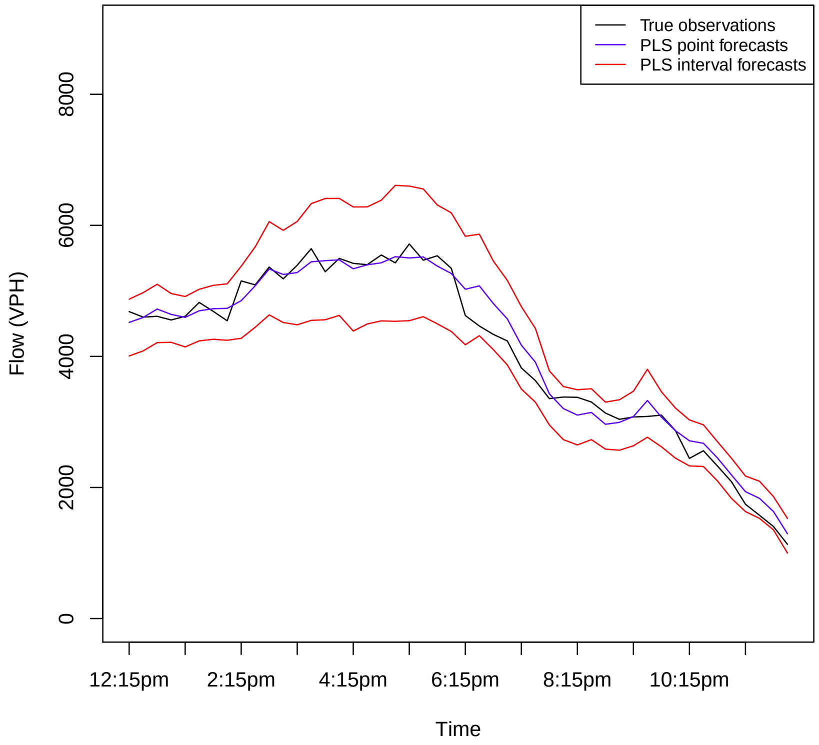


Figure 4.6. Partial forecast (noon-midnight) for March. 24th with 80% confidence intervals through BM and PLS Method.

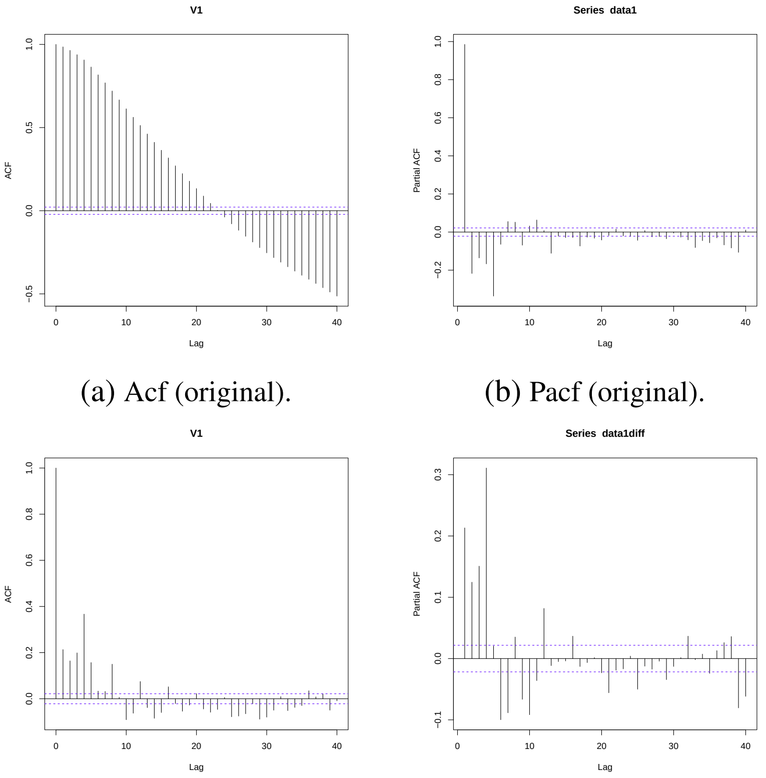
and Cook (1979), Nihan and Holmesland (1980) and Lee and Fambro (1999). ARIMA Models with its underlying linear model has outperformed many of the complex or non-linear proposed models in the application of traffic flow prediction. Examples of this can be found in SETAR (Self Excited Threshold Autoregressive) model has shown low traffic prediction accuracy compared to ARIMA Van Hinsbergen (2007). Similarly, in Sun and Liu (2011) a non-linear LSTAR (Logistic Smooth Transition Autoregressive) model to predict traffic flows is presented and found that ARIMA is more robust at forecasting a single step ahead. LSTAR (Logistic Smooth Transition Autoregressive) is superior in forecasting periods of low traffic, which is often not of great interest for traffic practitioners. In Chen et al. (2011), ordinary traffic flow prediction ARIMA is sufficient and outperformed the GARCH-ARIMA model. Examples of ARIMA being employed as a benchmark model can be found in Kamarianakis and Prastacos (2003), Smith et al. (2002) and Hamed and

Al-Masaeid (1995). ARIMA is therefore considered more appropriate for comparison analysis as it is a well established and researched model. In this study the details of the ARIMA model are omitted, the reader should refer to Brockwell and Davis (2002), which is a seminal book on traditional time series methodologies. In this study, it is being used as a benchmark for comparing the forecasting performance.

ARIMA Models can be non stationary; ARMA Models are stationary. Firstly, time series data is checked for stationarity. Autocorrelation function and partial autocorrelation function plots for the traffic flow data are shown in Figure 4.7a and 4.7b. Figure 4.7a clearly shows a relatively slow decay of correlation while Figure 4.7b indicates possibility of a unit root. First order difference is carried out and resultant results are illustrated in Figure 4.7c and 4.7d. The results depicts that first difference results in a stationary time series as the autocorrelation decay rapidly and is almost zero at a lag of 9, while a lag of 4 is found to be significant. In case of partial autocorrelation plot, again the fast decaying is clear and a lag of 4 appears to be significant. The purpose of differencing is to make time series data stationary, as the level of the series and the covariances stays roughly constant over time. However, to ensure that time series data does not have a unit root and is stationary after first difference , it is further cross checked by formally testing through series of tests (Augmented Dickey Fuller(*adf*), Kwiatkowski, Phillips, Schmidt, and Shin (*kpss*) and Phillips Perron (*pp*)) test at 95% confidence level. The "tseries" package in "R" is used to perform these tests. The "*adf*", and "*pp*" test are used to check the null hypothesis that differenced data set has a unit root against a stationary root alternative. Similarly "*kpss*" test is used with the null hypothesis that difference dataset has a stationary root against a unit-root alternative. The *p-values* of 0.01 in case of *adf* and *pp* test suggest to reject null hypothesis in favor of alternative hypothesis of stationarity at $\alpha = 0.05$. In case of *kpss* test the high *p-value* of 0.09997 suggest that we failed to reject the null hypothesis of stationarity with confidence level of $\alpha = 0.05$. The tests details are summarized in Table 4.2. As time series is stationary after first difference, the next step is to select the appropriate ARMA model, which means finding the most appropriate values of p and q for an ARMA(p, q) model. The package "forecast" in "R" is used to find the appropriate ARMA model for the time series basing on Akaike Information Criteria (AIC). An ARMA $(4, 4)$ is found appropriate, which corresponds to a model of fourth order, $p = 4$, and a moving average of 4, $q = 4$. One of the inherent limitations of ARIMA and other time series models, is essentially that

Table 4.2. Summary - Testing for stationarity of differenced data

Test	H_0	α	P-value	Result
ADF(*adf*)	unit root	0.05	0.01	Reject H_0
PP(*pp*)	unit root	0.05	0.01	Reject H_0
KPSS(*kpss*)	stationarity	0.05	0.09997	Fail to Reject H_0



(a) Acf (original).

(b) Pacf (original).

(c) Acf (diffrence).

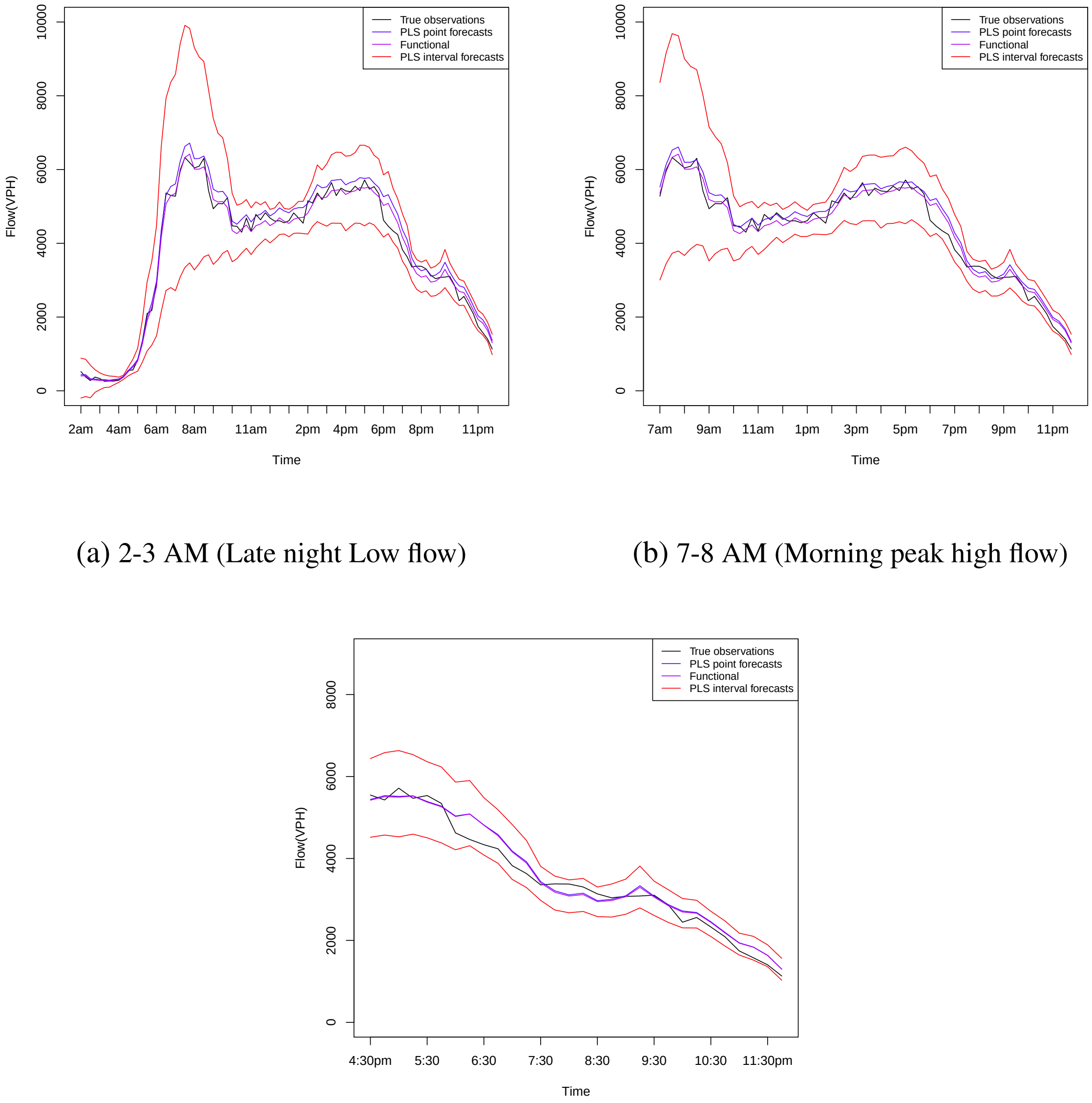
(d) Pacf (difference).

Figure 4.7. Autocorrelation and partial autocorrelation function plot of original and difference data.

they are "backward looking." Meaning the long-term forecast eventually converges to the sample mean. As the forecast horizon increases, the prediction interval gets wider and thus render any forecast greater than few steps practically meaningless. For this reason, only one hour forecast is made through ARIMA and compared with the first four forecast values of the complete 24 hours forecast resulting from the functional model.

4.4.3. Performance Comparison. The one step ahead forecast produced by functional method means complete next day forecast. Similarly, incase of partially observed traffic profiles, the functional context implies forecast for the remaining part of the day. However, only one hour ahead or four forecast points from the complete day or remainder day is considered in this study due to the reason that non functional methods such as ARIMA, when used in the context of traffic flow forecasting, is usable only for a short forecast horizon.

Results obtained from functional along with its dynamic updating variants and ARIMA are compared for 15-minutes interval data. The traffic flow profile fluctuates during a given day with extremely high peaks during morning and evening rush hours. The late night or early morning traffic periods exhibits very low traffic flows which is expected behavior of travelers. In order to fully evaluate the performance of the competing methods, three one hour periods are considered representing low-flow period, which occurs after midnight e.g. 2-3 AM, morning peak period from 7-8 AM and evening peak period from 4:30-5:30 PM is taken for the evening peak period. The results for the low flow period ($2-3$ AM) are summarized in Tables 4.3 and 4.4. Performance indicators shows the superiority of functional approach over ARIMA modeling in forecasting the period associated with low flow. In fact, functional method here implies complete forecast of next day rather updating the partially observed traffic profile through the PLS method. Functional method is followed by the functional (PLS) method in predictive accuracy. Although ARIMA method does not perform well overall, it is found to be better in predicting one step ahead flow values. With increase in forecast steps as well as sharp decrease in traffic flows, ARIMA's performance deteriorated and fails to forecast reasonably for the remaining steps ahead. In case of the high flow period for the morning (7-8 AM), in overall context, the functional method along with functional (PLS) again perform better than ARIMA, refer Tables 4.5 and 4.6. The functional method performs better in all the four steps ahead forecasting. In comparison to performance during low flow period, the performance of functional method improves with respect to functional (PLS) as MAPE gap increases with functional method achieving a MAPE of 1.530 in comparison with 4.268 for functional(PLS). The observed morning peak data increases sharply from step 1, 5276 vph to 5992 vph in step 2, which is very effectively and closely captured by functional method with the forecast value of 5896 vph. This reflects the method's ability in response to sharp increase in flow changes.



(a) 2-3 AM (Late night Low flow)

(b) 7-8 AM (Morning peak high flow)

(c) 4:30:5:30 PM (Evening peak high flow)

Figure 4.8. Partially observed profile forecast for low and high flow periods with 80% prediction intervals

The third part of result relates with the evening peak hour from 4:30-5:30 PM. Overall, functional (PLS) method performed slightly better than functional method with a MAPE of 2.136 in comparison with 2.172 achieved by functional method, for results refer Tables 4.7 and 4.8. For first and third steps ahead forecast, functional(PLS) methods achieves the

Table 4.3. Comparison four steps ahead prediction (Low Volume period, 2-3 AM)

Models	1 Step ahead	2 Step ahead	3 Step ahead	4 Step ahead
Actual	520	376	276	372
Functional	400	410	304	294
Functional(PLS)	434	438	326	320
ARIMA	500	528	521	506

Table 4.4. Performance measures four steps ahead prediction (Low Volume period, 2-3 AM)

Performance measures	RMSE	MAE	MAD	MAPE
Functional	74.914	65.148	83.337	15.865
Functional(PLS)	64.538	63.001	86.282	16.402
ARIMA	159.388	137.382	211.861	42.291

Table 4.5. Comparison four steps ahead prediction (High flow period,7-8 AM)

Models	1 Step ahead	2 Step ahead	3 Step ahead	4 Step ahead
Actual	5276	5992	6324	6188
Functional	5323	5896	6287	6377
Functional(PLS)	5528	6129	6526	6610
ARIMA	5692	6280	6723	6786

best results, while for the remaining two steps functional methods performs better. Again ARIMA performs the worst. The significant increase of flow from 5428 in step 1 to 5716 in step 2 is closely forecasted by functional(PLS) method. Figure 8 shows the forecast performance of all methods. It is important to note that forecast of partially observed profiles forecast is complete forecast for the remaining portion of the day, however only first four values are considered for comparison in all the three cases.

Results clearly indicates the superior forecasting performance of the functional approach over standard time series benchmark of ARIMA model. ARIMA only performs better in case of one step ahead forecast for low flow period. In all the remaining cases, its performance remain behind the functional and functional(PLS) method. The first perfor-

Table 4.6. Performance measures four steps ahead prediction (High flow period,7-8 AM)

Performance measures	RMSE	MAE	MAD	MAPE
Functional	109.877	92.090	105.552	1.530
Functional(PLS)	274.424	253.244	336.166	4.268
ARIMA	439.60	425.335	604.431	17.681

Table 4.7. Comparison four steps ahead prediction (High flow period, 4:30-5:30 PM)

Models	1 Step ahead	2 Step ahead	3 Step ahead	4 Step ahead
Actual	5548	5428	5716	5468
Functional	5389	5478	5460	5491
Functional(PLS)	5440	5534	5515	5530
ARIMA	5308	5228	5249	5249

Table 4.8. Performance measure four steps ahead prediction (High flow period, 4:30-5:30 PM)

Performance measures	RMSE	MAE	MAD	MAPE
Functional	153.080	122.024	154.94	2.172
Functional(PLS)	129.636	119.222	158.538	2.136
ARIMA	301.457	281.50	340.557	5.046

mance measure Root Mean Square Error(RMSE) represents the sample standard deviation of the differences between predicted values and observed values. ARIMA has highest RMSE value for PM peak period. The second performance measure Mean absolute Error(MAE) mean absolute error is an average of the absolute errors, mathematically, expressed as $e_i = |f_i - y_i|$, where f_i and y_i are forecasted and true values respectively. It evaluates forecast performance disregard of the direction of over-or under-prediction. It is interesting, that for low flow period, although functional method has lowest MAPE but has higher MAE value compared to functional (PLS) method. It implies that functional (PLS) has less error in absolute terms as compare to functional method for low flow period. The third performance measure Mean absolute Deviation (MAD) averages the absolute deviations and thus give less weight to the larger deviations. The MAD results are in line with the other two measures explained. The last measure Mean Absolute Percent Error (MAPE) also reflects the better performance of the functional and functional (PLS) methods. In overall context, functional method exhibits more accurate forecasts. This finding is significant for the short-term as well as medium-term traffic forecasting. It implies that, one can have a reasonably accurate forecast for the complete next day in advance under normal operating conditions. In traffic domain, the prime importance is method's ability to forecast the morning and evening peaks, where functional approach shows significantly accurate results as well as demonstrated its ability to forecast the sharp increase and decrease in the traffic flows.

4.5. CONCLUSION

This study is an effort to demonstrate the ability of functional approach in addressing a perpetual problem of traffic engineering "consistent and accurate traffic flow forecasting model". The functional model based on decomposition of functional principal components is proposed for short and medium-term traffic flow forecasting. The results demonstrate that functional model as well as its variant (PLS) method for dynamic updating forecast offer significantly improved forecasting performance in comparison to conventional time series approach, represented herein by ARIMA. For this purpose, one hour ahead forecast comparison was made of all the competing models. The empirical results demonstrate that functional and functional(PLS) methods accurately estimates the low volume or free flow as well as high flow periods. It demonstrates that the functional approach has the ability to characterize cyclical dynamics of short-term traffic forecast and thus provide better forecast performance than an ARIMA model. In addition, the ability of functional method to forecast complete day ahead with consistency is significant.

A major problem in traffic management to forecast in real time is addressed by employing PLS technique to dynamically update the partially observed traffic profile. However, its interesting that results show that functional method seems to be slightly better than dynamic updating (PLS) method. The four steps (15 minutes each) out of one day ahead forecast obtained through functional method is accurate as well as consistent in forecasting the traffic flows for this empirical study. The consistent and accurate traffic flow forecast encourages to apply the functional approach in forecasting the other variables of traffic parameters like speed and travel times.

The functional approaches provide a useful means to produce not only short-term but also medium-term forecasts. That is, the functional approach results in a full day's traffic profile. Whether it is complete day ahead or predicting the rest of the day given a partial profile. The implications of such models would do well to serve as useful for the short-term management of traffic as well as providing information for planning of routes for users. Furthermore, functional methods provide useful information for ITS, ATIS, and ATM systems.

The future direction of this work is aimed towards the development of a routing methodology, which makes use of the daily traffic prediction to develop user routes in a given traffic network that takes into account traffic at the time of travel for short or medium time horizons.

5. CONCLUSIONS

This dissertation presents three different methodologies with a focus on improving the existing methods available for traffic pattern analysis and forecasting. All the previous research efforts for traffic flow pattern analysis are aimed at developing a forecast model. Which, requires a thorough understanding of traffic behavior often derived from analyzing historical data. This research also focussed on the same goal but adopted consistent and robust methods in exploring the traffic data as well as proposing an accurate forecasting model. The dissertation investigates how a shape based classification provides insights into traveler's behavior demonstrated through daily traffic flow profiles observed on freeways. It also discusses the advantages gained by using a shape based classification technique. The dissertation used an artificial intelligent technique (BPNN) to validate the results obtained from shape classification. The final goal of an accurate and consistent forecasting model is achieved by proposing a functional model based on FPC's decomposition.

In the first study, a data exploration is carried out by studying the traffic flow over a period of 5 years. This research effort is the first attempt to classify the traffic flow profiles solely based on their shape characteristics. All previous efforts on traffic flow pattern analysis have employed techniques that are non-shape based. These efforts also used less robust clustering methods while some employed techniques from contemporary fields that are often customized for some other purpose. The proposed methodology addressed the prevailing adhocism on the subject. It provides practitioners a logical and consistent method to explore and classify the traffic flow profiles by incorporating their shape features. The proposed methodology simplifies the pattern analysis as shape covers all features comprehensively. Hence, features like total flows, peak flows, time of the peak flows and off-peak flows are not required to be analyzed separately as shape incorporates all of these key characteristics. Similarly, no pre-classification is required in the proposed methodology, a departure from some existing methodologies where two step clustering is performed. The use of partition around mediods (PAM) algorithm is another feature that gives added strength to the methodology. As a result tight, stable, and statistically significant clusters are obtained. The demonstrated capability of the GSD function in detecting anomalies alongside classifying separately the days with influence of weather and social events is also found promising. However, it is also observed that once translating the traffic flow profiles into their respective GSDs, a shape based cumulative distribution is obtained but in the process the corresponding time domain is lost. This limitation of the research is significant in case the obtained results are required to be translated back to generate the simulated profiles e.g. simulation etc.

In order to validate the shape based classification, a methodology based on Back Propagation Neural Network (BPNN) was employed in second study. This was an important step before moving to the ultimate goal of developing a forecasting model. A comparative classification analysis of original and GSD transformed data reveals that shape based classification is significantly more stable and consistent in comparison to non-shaped based classification, techniques, and processes. In fact, these results demonstrate the necessity of considering functional shape in traffic flow analysis. In addition, the results reinforce the need for clustering prior to prediction. It is also determined that a span of two through four years of traffic data is found sufficient for training to produce satisfactory BPNN performance. This provides a fair guideline about data usage for the employment of Neural Network (NN) in traffic research. The BPNN methodology provides a simple method to compare the classification results obtained through different approaches and can be applied for future classifications comparisons in traffic domain.

The contribution of the third study is using functional data analysis techniques applied to the traffic domain. This brings a change in philosophy towards the handling of traffic time series data and provides a motivation to develop functional time series for traffic forecasting. The proposed functional approach provides a useful means to produce not only short-term but also medium-term forecasts. That is, the functional approach results in a full day's traffic profile. Whether it is complete day ahead or predicting the rest of the day from given a partially observed profile. It demonstrates that the functional approach has the ability to characterize cyclical dynamics of short-term traffic forecast. Thus, this method provides better forecast performance than the well established ARIMA model. The implications of such models would serve useful for the short-term traffic management as well as providing information for planning of routes for users. Furthermore, functional methods provide useful information for ITS, ATIS, and ATM systems. The consistent and accurate traffic flow forecast encourages to apply the functional approach in forecasting the other variables of traffic parameters like speed and travel times.

6. FUTURE WORK

This dissertation opened a number of avenues for future research. These research avenues are summarized below:

1: The research has introduced shape based classification that demonstrated promising results in terms of stable, tight, and statistically significant grouping. However, in the process the time domain is lost and only cumulative distribution is available describing the shape features of the daily traffic flow profile. Relating the obtained cumulative distribution back to time domain (24 hours of a day) is not possible. It is a drawback for example in case a simulation is required to generate a traffic profile from obtained distribution. This aspect is opened for future research and some technique i.e. (mapping etc) may be explored to address this issue.

2: Real time traffic forecast is a major component of any ITS System. This research has addressed this issue by exploring the PLS technique to dynamically update the partially observed traffic flow profile. There are number of other regression and non regression based techniques like Ridge Regression (RR) and Block Moving (BM) being already used in the contemporary fields. These techniques can be employed in traffic domain and their relative forecast performance be compared with PLS technique.

3: The proposed methodology is applied only for flow parameter. It will be interesting to observe its performance once applied to other traffic parameters like speed profiles, space/time headways, and density/occupancy etc. This application may include both aspects of exploring the patterns as well as functional forecasting.

4: The proposed research is based on a single location (single detector location) on a freeway. It implies that at present the demonstrated results are valid for specific traffic and weather conditions. To observe relative performance, the proposed methodology may be applied at multiple locations in future.

5: It is well established in the literature that no single technique is perfect in non-functional traffic flow forecasting. Researchers in the past experimented by augmenting two different techniques from divergent fields like time series model and NN from Artificial Intelligence (AI), to improve forecast performance. Development of a hybrid functional model by using BPNN and functional time series model for improving forecast accuracy can be an interesting future research.

6: Development of a routing methodology based on proposed functional forecasting model is yet another future research direction. It makes use of the daily traffic prediction to develop user routes in a given traffic network and also takes into account traffic at the time of travel for short or medium time horizons.

APPENDIX

PARTITION AROUND MEDIODS (PAM)/K-MEDOIDS ALGORITHM

The goal of the PAM algorithm is to find a sequence of objects called *medoids* that are centrally located in clusters. Hence, it can be summarized that the algorithm seeks to minimize the average dissimilarity of objects to their closest selected object. The algorithm has two phases Kaufman and Rousseeuw (2009):

i The first phase is referred to as the *build*, a collection of k objects are selected for an initial set S.

ii The second phase is called *swap*, where one tries to improve the quality of the clustering by exchanging selected objects with unselected objects.

The PAM algorithm is a well established and known algorithm. Due to this popularity the details are omitted in this research. Refer to Kaufman and Rousseeuw (2009) for detailed description of this algorithm. The following summary of the algorithm is provided to assist the reader:

```
1: Select $k$ representative GSD vectors arbitrarily to represent medoids.
2: For each pair of non-selected GSDs $h$ and selected GSD $i$, calculate the total swapping cost $TC_{ih} = \sum_i C_{jih}$, where $C_{jih} = d(j,h) - d(j,i)$.
3: For each pair of $i$ and $h$,
     – If $TC_{ih} < 0$, $i$ is replaced by $h$
     – Then assign each non-selected GSD to the most similar representative object.
4: Repeat steps 2 and 3 until the change in clusters is minimal or no change.
```

BIBLIOGRAPHY

Aguilera, A.M., Ocana, F.A., and Valderrama, M.J., “ Forecasting time series by functional PCA. Discussion of several weighted approaches”. Computational Statistics 14.3 (1999): 443–467.

Ahmed, M.S., and Cook, A.R., “Analysis of freeway traffic time-series data by using Box-Jenkins Techniques”. Transportation Research Record 722 (1979): 1–9.

Ahmed, M. S., “Stochastic processes in freeway traffic Part I. Robust prediction models”. Traffic Engineering and Control 24. HS-035 775 (1983).

Antoch, J., “Functional linear regression with functional response: application to prediction of electricity consumption”. Functional and Operatorial Statistics. Physica-Verlag HD, 2008. 23–29.

Ballarin, V. Y., and Valentinuzzi, M., “Segmentacin en imgenes de Resonancia Magntica de Cerebro utilizando Morfologa Matemtica”. Taf del Valle (September 2001). (Publicadas en CD), 2001.

Box, G. E. P., and Jenkins, G. M., “Time series analysis: Forecasting and control (revised edition).” Holden Day, San Francisco (1976).

Boxill, Sharon Adams, and Lei Yu., “An Evaluation of Traffic Simulation Models for Supporting ITS”. Houston, 2000.

Brockwell, P.J., “Introduction to time series and forecasting.” Vol. 1. Taylor & Francis, 2002.

Cetiner, B.G., Sari, M., and Borat, O., “A neural network based traffic-flow prediction model”. Mathematical and Computational Applications 15.2 (2010): 269–278.

Chang, L.-Y., “Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network”. Safety Science 43 (2005): 541–557.

Chen, C., Hu, J., Meng, Q., and Zhang, Y., “Short-time traffic flow prediction with ARIMA-GARCH model.” In Intelligent Vehicles Symposium (IEEE) (IV). IEEE (2011).

Chen, H. and Grant-Muller, S. “Use of sequential learning for short-term traffic flow forecasting”. Transportation Research Part C: Emerging Technologies 9 (2001): 319–336.

Cheslow, M., Hatcher, S.G., and Patel, V.M., “An initial evaluation of alternative intelligent vehicle highway systems architectures”. No. MTR 92W0000063. 1992.

Chrobok, R., Kaumann, O.,Wahle, J., and Schreckenberg,M. Different methods of traffic forecast based on real data. European Journal of Operational Research, 155(3), 558568.

Chiou, J.M., “Dynamical functional prediction classification with application to traffic flow prediction”. The Annals of Applied Statistics 6.4 (2012): 1588–1614.

Chung, E. and Rosalion, N. Short term traffic flow prediction. Australasian Transport Research Forum (ATRF) , 24th, 2001, Hobart, Tasmania Australia.

Chung,E., “Classification of traffic Pattern”. Proceedings of the 11th world Congress on ITS (2003): 687–694.

Dia, H., “An object-oriented neural network approach to short-term traffic forecasting”. European Journal of Operational Research 131.2 (2001): 253–261.

Dijk, D. Terasvirta, T., and Franses, P.H., “Smooth transition autoregressive models - a survey of recent developments, Econometric Reviews”. Taylor and Francis Journals 21.1 (2002): 1–47.

Donoho, D.L., ”High-dimensional data analysis: The curses and blessings of dimensionality.” AMS Math Challenges Lecture (2000): 1-32.

Erbas, B., Hyndman, R.J., and Gertig, D.M., “Forecasting age-specific breast cancer mortality using functional data models”. Statistics in Medicine 26.2 (2007) : 458–470.

Erlingsson, S., Jonsdottir, A. M., and Thorsteinsson, T., “Traffic stream modelling of road facilities”. Transport Research Arena Europe (2006).

Gaston-Romeo, M., Leon, T., Mallor, F., and Ramrez-Santigosa, L., “A morphological clustering method for daily solar radiation curves”. Solar Energy, 85.9 (2011): 1824–1836.

Goh, A.T.C, ”Back-propagation neural networks for modeling complex systems.” Artificial Intelligence in Engineering, 9(3): (1995), 143–151.

Gonzlez, R.Y., and Woods, R., “Tratamiento Digital de imgenes”. Addison Wesley, 1996.

Green, P.J., and Silverman, B.W., “Nonparametric regression and generalized linear models: A roughness penalty approach”. CRC Press, 1993.

Guardiola, I. G., and Mallor, F., “A nonparametric method for detecting unintended electromagnetic emissions”. Electromagnetic Compatibility, IEEE Transactions on 55.1 (2013): 58–65.

Guardiola, I. G., Leon, T., and Mallor, F., ”A functional approach to monitor and recognize patterns of daily traffic profiles.” Transportation Research Part B: Methodological 65 (2014): 119-136.

Guardiola, I. G., Wasim, I. and Samaranayke,V.A. “On traffic flow pattern shape classification and analysis”. Manuscript submitted for Publication.

Hamed, M.M., and Al-Masaeid, H.R., “Short-term prediction of traffic volume in urban arterials”. Journal of Transportation Engineering 121.3(1995): 249–254.

Hashemi, R.R., Le blanc, L.A., Rucks, C.T., Shearry, A., “A neural network for transportation safety modeling”. Expert Systems with Applications 9.3 (1995): 247–256.

Heaton, J. “Introduction to neural network for Java”. 2nd Edition, ISBN 1604390085.

Hogberg, P., “Estimation of parameters in models for traffic prediction: a non-linear regression approach”. Proceedings of the 11th world Congress on ITS (2003): 687–694.

Hormann, S., and Kokoszka, P., “Weakly dependent functional data”. The Annals of Statistics 38.3 (2010): 1845–1884.

Horvath, L., Huskova, M., and Kokoszka, P., “Testing the stability of the functional autoregressive process”. Journal of Multivariate Analysis 101.2 (2010): 352–367.

Horvth, L., and Kokoszka, P. “Inference for functional data with applications”. Vol. 200. Springer Science and Business Media, 2012.

Hyndman, R.J., and Shang, H.L., “Forecasting functional time series (with discussion)”. Journal of the Korean Statistical Society 38.3 (2009): 199–221.

Hyndman, R.J., and Shang, H.L., “Rainbow plots, bagplots, and boxplots for functional data”. Journal of Computational and Graphical Statistics 19.1 (2010): 29–45.

Ivan, J.N., and Sethi, V., “Data fusion of fixed detection and probe vehicle data for incident detection”. Computer-aided Civil and Infrastructure Engineering, 13.5 : 329–337.

Jiang, Xiaomo, and Hojjat Adeli. ”Dynamic wavelet neural network model for traffic flow forecasting”. Journal of Transportation Engineering 131.10 (2005): 771–779.

Kamarianakis, Y., and Prastacos, P.,“Forecasting traffic flow conditions in an urban network: comparison of multivariate and univariate approaches”. Transportation Research Record 1857.1 (2003): 74–84.

Kamarianakis, Y., Shen, W., and Wynter, L., “Real-time road traffic forecasting using regime-switching spacetime models and adaptive LASSO (with discussion)”. Applied Stochastic Models in Business and Industry 28.4 (2012): 297–315.

Kargin, V., and Onatski, A., ”Curve forecasting by functional autoregression.” Journal of Multivariate Analysis 99.10 (2008): 2508–2526.

Karim, A., and Adeli, H., Incident detection algorithm using wavelet energy representation of traffic patterns. Journal of Transportation Engineering, 128.3: 232242.

Kaufman, L. and Rousseeuw, P. J., “Finding groups in data: an introduction to cluster analysis”. Vol. 344. Wiley. com.

Khan, S. I., and Ritchie, S. G., “Statistical and neural classifiers to detect traffic operational problems on urban arterials”. Transportation Research Part C: Emerging Technologies 6.5 (1998): 291–314.

Kidzinski, L., ”Functional Time Series.” CHILI Laboratory, Ecole polytechnique federale de Lausanne, RLC D1 740, CH-1015, Lausanne, Switzerland, arXiv preprint arXiv:1502.07113. (2015): 1–12.

Kirby, H. R., Waston, S. M., and Dougherty, M. S., “Should we use neural networks or statistical models for short-term motorway traffic forecasting?”. Int. J. Fore-casting 13 (1997): 43–50.

Lan, Lawrence W., Jiuh-Biing Sheu, and Yi-San Huang. ”Investigation of temporal freeway traffic patterns in reconstructed state spaces”. Transportation Research Part C: Emerging Technologies 16.1 (2008): 116-136.

Lee, S., and Fambro, D.B., “Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting”. Journal of the Transportation Research Board 1678.1 (1999): 179–188.

Levitin, D. J., “Introduction to Functional Data Analysis”. Canadian Psychology Vol 48.3 (2007): 135-155.

Levin, M., and Tsao, Y.D., “On forecasting freeway occupancies and volumes”. Transportation Research Record 773 (1980): 47–49.

Leutzbach, W., “Introduction to the Theory of Traffic Flow”. Springer-Verlag (1998), New York.

Lozano, A., Manfredi, G., and Nieddu, L. (2009). An algorithm for the recognition of levels of congestion in road traffic problems. Mathematics and Computers in Simulation, 79(6), 19261934.

Nicholson, H., and Swann, C., The prediction of traffic flow volumes based on spectral analysis. Transportation Research, 8.6 (1974): 533–538.

Nihan, N.L., and Holmesland, K.O., “Use of the box and Jenkins time series technique in traffic forecasting,” Transportation vol. 9.2 (1980).

Okutani, I., and Stephanedes, Y. J., “Dynamic prediction of traffic volume through kalman filtering theory”. Transportation Research Part B: Methodological, 18.1 (1984): 1–11.

Pant, P.D., Balakrishnan, P., “Neural network for gap acceptance at stop-controlled intersections”. Journal of Transportation Engineering 120.3 (1994): 432–446.

Patra, P. K., Vipsita, S., Mohapatra, S., and Dash, S. K., “A novel approach for pattern recognition”. International Journal of Computer Applications: 9 8 (2010): 19–23.

Rakha, H., and Van Aerde, M., “Statistical analysis of day-to-day variations in real time flow data”. Transportation Research Record (1995): 26–34.

Ramsey, J.O., and Silverman, B.W., “Functional data analysis”. 2nd edition. Springer-Verlag, Newyork.

Sadek, S., Al-Hamadi, A., Michaelis, B., and Sayed, U., A statistical framework for real time traffic accident recognition. J. Signal and Information Processing, 1.1 (2010): 77–81.

Serra, J., “Image Analysis and Mathematical Morphology”. Academic Press, 1982.

Shang, H.L., and Hyndman, R.J.,“Nonparametric time series forecasting with dynamic updating”. Mathematics and Computers in Simulation 81.7 (2011): 1310–1324.

Shang, H.L., “ftsa: An R Package for Analyzing Functional Time Series”. The R JOURNAL 5.1 (2013): 64–72.

Shen, H., and Huang, J.Z., “Interday forecasting and intraday updating of call center arrivals”. Manufacturing and Service Operations Management 10.3 (2008): 391–410.

Smith, B.L., Williams, B.M., and Oswald, R.K., “Comparison of parametric and nonparametric models for traffic flow forecasting”. Transportation Research Part C:Emerging Technologies 10.4 (2002): 303–321.

Stathopoulos, A., and Karlaftis, M.G., ”A multivariate state space approach for urban traffic flow modeling and prediction”. Transportation Research Part C: Emerging Technologies 11.2 (2003): 121–135.

Stephanedes, Y., Kwon, E., and Michalopoulos, P., “On-line diversion prediction for dynamic control and vehicle guidance in freeway corridors”. Number 1287 (1990).

Sommer, C., and Falko, D., “Progressing toward realistic mobility models in VANET simulations”. Communications Magazine, IEEE 46.11 (2008): 132–137.

Soriguera, F., ”Deriving traffic flow patterns from historical data”. Journal of Transportation Engineering 138.12 (2012): 1430–1441.

Sun, H., Liu, H.X., Xio, H., and He, R.R., “Use of local linear regression model for short-term traffic forecasting”. Transportation Research Record 1836.1 (2003): 143–150.

Sun, X., Liu, T., “A Star model for urban short-term traffic flow forecasting”. Advanced Forum on Transportation of China (AFTC 2011), 7th. IET, 2011.

Tan, M. C., Wong, S. C., Xu, J. M., Guan, Z. R., and Zhang, P., ”An aggregation approach to short-term traffic flow prediction”. Intelligent Transportation Systems, IEEE Transactions 10.1 (2009): 60–69.

Terasvirta, T., and Anderson, H.M., “Characterizing nonlinearities in business cycles using smooth transition autoregressive models”. Journal of Applied Econometrics 7.S1 (1992): S119–S136.

Terasvirta, T., “Specification, estimation, and evaluation of smooth transition autoregressive models”. Journal of the American Statistical Association 89.425 (1994): 208–218.

Van Hinsbergen, J.W.C., and Sanders, F.M., “Short-term Prediction models”. (2007).

Varaiya, P., “What weve learned about highway congestion”. ACCESS Magazine 1.27 (2005).

Vincent L., and Dougherty E., “ Morfological Segmentation for Textures and Particles”. Digital Image Processing Methods, Rochester, New York, 1994.

Wei, Chien-Hung, and Ying Lee. “Sequential forecast of incident duration using Artificial Neural Network models”. Accident Analysis and Prevention 39.5 (2007): 944–954.

Weijermars, W. A. M., and Van Berkum, E. C., “Analysis of highway flow patterns using cluster analysis”. Intelligent Transportation Systems, Proceedings (2005): 308–313.

Weijermars, W. A. M., “Analysis of urban traffic Patterns using clustering”. University of Twente, 2007.

Weil, R., Wootton, J., and Garcia-Ortiz, A., ”Traffic incident detection: Sensors and algorithms”. Mathematical and computer modelling 27.9 (1998): 257–291.

Wild, D., “Short-term forecasting based on a transformation and classification of traffic volume time series”. International Journal of Forecasting 13.1(1997): 63–72.

Williams, B.M., Durvasula, P.K., and Brown, D.E., “Urban Freeway Traffic Flow Prediction: Application of Seasonal Autoregressive Integrated Moving Average and Exponential Smoothing Models”. Transportation Research Record: Journal of the Transportation Research Board 1644.1 (1998): 132–141.

Williams, B.M., “Multivariate vehicular traffic flow prediction: Evaluation of ARIMAX modelling”. Transportation Research Record: Journal of the Transportation Research Board 1776.1 (2001): 194–200.

Williams, B.M., and Hoel, L.A., “Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process:Theoretical basis and empirical results”. Journal of transportation engineering 129.6 (2003): 664–672.

Xie, Y., Zhang, Y., and Ye, Z., “Short-term traffic volume forecasting using Kalman filter with discrete wavelet decomposition”. Computer?Aided Civil and Infrastructure Engineering 22.5 (2007): 326–334.

Yasdi, R., “Prediction of road traffic using a neural network approach”. Neural computing and applications 8.2 (1999): 135–142.

Yin, H., Wong, S. C., Xu, J. and Wong, C. K. “Urban traffic flow prediction using fuzzy neural approach”. Transportation Research Part C: Emerging Technologies 10.2 (2002): 85–98.

Yu, C., and Lam, K.C., ”Applying multiple kernel learning and support vector machine for solving the multicriteria and nonlinearity problems of traffic flow prediction.” Journal of Advanced Transportation 48.3 (2014): 250-271.

Zhang, Y., and Ye, Z., “Short-term traffic flow forecasting using fuzzy logic system methods”. Journal of Intelligent Transportation Systems 12.3 (2008): 102–112.

Zheng, W., Lee D.H., and Shi Q., “Short-term freeway traffic prediction: Bayesian combined neural network approach”. Journal of transportation engineering 132.2 (2006): 114–121.

Zhou, X.W. “A smooth transition autoregressive model for electricity prices of Sweden”. Master thesis of statistics, Department of Statistics, Hogskolan Dalarna (2009).

Zhu, W., and Barth, M., (2006). Vehicle trajectory-based road type and congestion recognition using wavelet analysis. Intelligent Transportation Systems Conference, 2006. ITSC06. IEEE. IEEE, 2006.

Zhu, Z., and Chun, Y., ”Visco-elastic traffic flow model.” Journal of Advanced Transportation 47.7 (2013): 635-649.

VITA

Wasim Kayani was commissioned in Pakistan Army Corps of Engineers in 1989. He has a Bachelor of Science in (Physics,Double Maths)from University of Peshawar, Pakistan as well as in (Civil Engineering) from Military College of Engineering, Risalpur, Pakistan. He obtained a Master of Science in War studies in 2003 from Pakistan, and Master of Science in Civil Engineering from the Missouri University of Science and Technology USA, awarded in May 2011. He has earned his Doctor of Philosophy in Civil Engineering with emphasis on Transportation Engineering from the same university in August 2015.