The Prediction of Traffic Flow with Regression Analysis

Ishteaque Alam
ID: 012 172 013

Department of Computer Science and Engineering

United International University

A thesis submitted for the degree of

MSc in Computer Science & Engineering

June 2018

Declaration

I, Ishteaque Alam, declare that this thesis titled, "The Prediction of Traffic Flow with Regression Analysis" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a MSc degree at United International University.
- Where any part of this thesis has never previously been submitted for a degree or any other qualification at United International University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ________________________

Date: ________________________

Ishteaque Alam

Certificate

I do hereby declare that the research works embodied in this thesis entitled "The Prediction of Traffic Flow with Regression Analysis" is the outcome of an original work carried out by Ishteaque Alam under my supervision.

I further certify that the dissertation meets the requirements and the standard for the degree of MSc in Computer Science and Engineering.

Signed: ________________________________

Date: ________________________________

Dr. Dewan Md. Farid

Associate Professor,
Department of Computer Science and Engineering,
United International University,
Dhaka-1212, Bangladesh.

Abstract

Road traffic management is necessary for smart cities. Lately, Intelligent Transport System (ITS) has become an important area of research to solve different road traffic related issues for making smart decision. It links people, roads and vehicles together using communication technologies to increase safety and mobility. Moreover, accurate prediction of road traffic is also important to manage traffic congestion. In this thesis work, we give a brief survey of recent researches on road traffic prediction and propose an innovative approach to estimate road traffic flow using regression analysis for the roads of Porto city (capital of Portugal). The 22 inductive loop traffic detectors (loop sensors) were used to collect traffic data around VCI (Via de Cintura Interna) motorway. Initially, we have applied data preprocessing and feature selection techniques on the traffic data and then applied several regression models (Linear Regression, SMO Regression, Multilayer Perceptron, M5P Regression Tree and Random Forest) to predict future traffic flow. Finally, we apply regression models including an ensemble model proposed in the study to predict road traffic flow in long-term based on historical traffic data and compare their performance. The experiment results show that prediction generated by M5P based regression tree tend to get the closest to actual traffic flow for different roads of Porto city, Portugal.

Published Papers

Work related to the research presented in this thesis has been published by the author in the following peer-reviewed conferences:

1. Ishteaque Alam, Dewan Md. Farid, and Rosaldo J. F. Rossetti, The Prediction of Traffic Flow with Regression Analysis, International Conference on Emerging Technology in Data Mining and Information Security (IEMIS 2018), February 23-25, 2018, Kolkata, India, pp. 1-10.
2. Ishteaque Alam, Dewan Md. Farid, and Rosaldo J. F. Rossetti, The Prediction of Traffic Flow with Regression Analysis, Advances in Intelligent Systems and Computing, Vol. 813, Springer, A. Abraham et al. (eds.), Emerging Technologies in Data Mining and Information Security, https://doi.org/10.1007/978-981-13-1498-8_58 (In Press)

Acknowledgements

I would like to thank my thesis supervisor Dr. Dewan Md. Farid, Associate Professor, Department of Computer Science and Engineering, United International University, Bangladesh. I feel grateful to him and wish my profound indebtedness to my honorable supervisor. Deep knowledge and keen interest of my supervisor in the field of Data Mining influenced me to carry out this thesis.

Also, I would like to thank our MSCSE coordinator Prof. Dr. Mohammad Nurul Huda sir for all of his suggestion and guidance throughout my masters study. And I am grateful to Prof. Dr. Khondaker Abdullah Al Mamun for allowing me to work with him as a graduate/research assistant at AIMS Lab, UIU.

Special thanks to Dr. Swakkhar Shatabda, Prof. Dr. Hasan Sarwar, Prof. Dr. Salekul Islam and Prof. Dr. Chowdhury Mofizur Rahman for all their teaching and advises which motivated and helped me a lot to build myself.

We would like to thank Dr. Rosaldo J. F. Rossetti, Assistant Professor, University of Porto, Portugal for his valuable advice and support and Armis (http://www.armis.pt/) for sharing the traffic data.

Finally, I would like to thank my parents for their unconditional love and support. Their love provided me inspiration and was my driving force that has pushed me to complete this dissertation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Traffic management has been a big issue since the beginning of road transportation, specially, in the busy cities. In present time, Los Angeles, New York, Moscow, San Francisco, Beijing and London are some of the most traffic congested cities around the world. Our Dhaka city is also struggling with the problem of traffic congestion, which, is literally frustrating for the people of the city. The increased number of vehicles on the road also adversely affect people's health. There are several tools and techniques have been used to solve this problem, such as, monitoring the traffic with modern techniques or using traditional light controls but, nothing seems good enough. It is required more research into this area to come up with better solutions to mitigate losses.

1.1 Motivation

The problem of traffic congestion affects the world's economy, environment and people's life. It also increases the stress level of people by making travel delays. Not only that, the stopping and initiating cars back again burns fuel at a greater level. This creates air pollution and thus, increases global worming. This fuel consumption also costs a lot of money. Proper investigation and monitoring are must to make smooth traffic flow. Intelligent Transport System (ITS) over the last decade, have been dealing with the management of road traffic flow to make the travel safer and more comfortable. ITS involves researchers to work with road traffic data. The information extracted from traffic data are obtained by using different sensors, such as, Pneumatic Road tube counting, Video Vehicle Detection, Piezoelectric Sensor, Magnetic Sensor etc. These sensors get placed on the road along roadside and constantly observe roads and store the information into database. So, the database gets quite a huge volume and thus, involves the researchers to deal with big data. In our previous work, we built a traffic data visualization tool (TDO) that involved real world road traffic Big data [1].

The road traffic data can be used to understand the behavior of vehicles to manage the road traffic. These data can also be used to predict traffic flow for different roads in a city. The prediction of traffic flow can be a real help to manage the road traffic. An amount of research have done to predict the traffic flow in short-term. But, only a handful of researches are there regarding long-term traffic flow prediction. Nevertheless, many researches and approaches for observing and predicting road traffic are required because, the situation and environment are different in different parts of the world. The involvement of big data makes this research area quite rich and creates opportunity for extensive future research. Making prediction for long-term traffic flow is going to have massive impacts by saving people's time and money.

1.1.1 Intelligent Transportation System (ITS)

ITS provides smart ways to improve traffic condition with the help of intelligent transportation technologies, such as, navigating cars, controlling traffic signals, recognizing number plates of cars, speed or CCTV systems etc. For efficient management of traffic, man have been providing solutions for decades to prevent commuters driving in the wrong direction to avoid congestion [2]. In recent times, ITS opened a gateway to find solutions to this problems using the cutting edge technologies. Floating car data (FCD) and GPS trajectory data has became an important data source for traffic data mining, which enables researchers to understand the traffic patterns [3], [4]. These understandings are important for better traffic management. Data scientists of modern time use these techniques to apply different machine learning methods to improve human life, such as, reducing road noises [5], forecasting the traffic demand [6] and even prediction of road accidents [7]. Also, prediction of road traffic congestion has become an emergent part in the field of ITS [8]. These predictions can be used later on for efficient route planning to help people plan their trip [9]. It is very clear that approaches of ITS through cutting edge technologies able transportation engineering to produce better mobility and safety. Therefore, ITS can make difference in saving lives and creating transportation infrastructures and technologies for the better future.

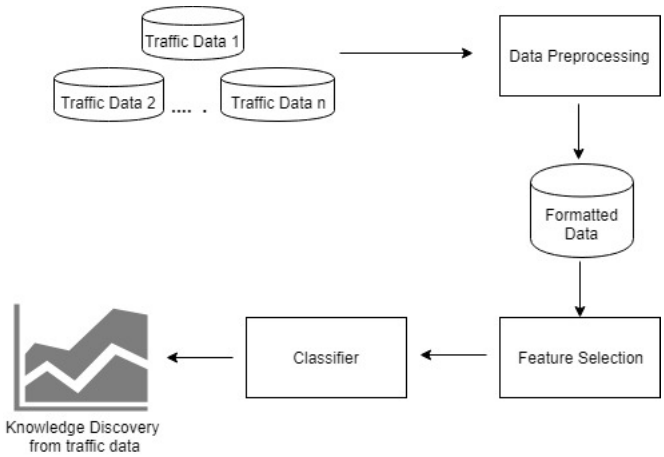


Figure 1.1: Traffic Data Mining - Extracting knowledge from the road traffic.

1.1.2 Mining Traffic Big Data

Big data is the new architectural shift of the modern world's enterprise computing. The massive increment of computed data everyday from different sources is the reason behind the origin of big data. These different sources include, social media data, retailer databases, financial data, health data, data generated from IoT (Internet of Things) etc. A solid big data platform makes it faster and easier for people to create unique data intensive products and services to improve people's lives. In most cases, parallel processing technique (Hadoop MapReduce) is used for handling or manipulating big data. The characteristics of big data is often referred to as 3V's or 5V's:

- Volume
- Velocity
- Variety
- Veracity
- Value

Traffic data mining in most of the cases involves dealing with big data. Because, it requires high volume of historical data to analyze the overall traffic pattern. And even for the prediction of traffic flow, it is better to gather a higher volume of previous road data for better prediction in short-term and in long-term [10]. The dataset that is used in this study was also a big data. It was a continuous observation of the roads of a city and thus, created enormous volume of data which was complex to analyze. Variety in data types were also present in the dataset. The details of our dataset is mentioned in Chapter 3.

1.2 Objectives of the Thesis

ITS related researches enable researchers to use road traffic data for forecasting road traffic flow. This is very necessary for cities with busy traffic. Our work aims to contribute in this field of ITS for making smart cities. This work studies the field of traffic data mining in recent times, and the recent approaches for prediction of road traffic. This work also presents an innovative approach of using regression models for long-term traffic flow forecasting. And also the study introduce an ensemble regression method and compare the performance of different regression models with a real world traffic data. The real world dataset was collected from Porto, Portugal. We applied the same regression models on different roads of the city to accurately identify the best performed regression model in terms of long-term traffic flow prediction. The major objectives of this thesis are summarized below:

1. Visualize and analyze the traffic big data.
2. Find the traffic pattern from historical traffic data.
3. Apply the data preprocessing and feature selection techniques on the traffic data.
4. Apply regression analysis methods (e.g. Linear Regression, SMO Regression, Multilayer Perceptron, and M5P regression tree) and compare their performance real time traffic data.

1.3 Thesis Contributions

The contributions of this thesis are summarized as follows;

- Recent works regarding the road traffic prediction and traffic data mining in the field of ITS have been reviewed with an organized manner.

- The study presented a new approach for forecasting the road traffic in long-term using a real world traffic dataset.
- Applied regression analysis for accurate traffic flow prediction.
- Proposed an ensemble approach combined with three regression models to predict traffic flow.
- This work also compared the performance of different regression models including the proposed approach and identified the best regression model to forecast road traffic for the city of Porto, Portugal.

1.4 Organization of the Thesis

The thesis is organized as follows:

Chapter 2 provides related works.

Chapter 3 presents the traffic data that was collected from city of Porto, Portugal.

Chapter 4 Discusses several regression models

Chapter 5 presents the experimental analysis.

Chapter 6 presents conclusion and future works.

Chapter 2

Related Work

Traffic data mining in the domain of ITS is providing the fundamental platform for managing and analyzing road traffic information to improve decision making. This enable researchers to propose solution for cities and communities to optimize their performance. Many studies focus to train different machine learning algorithms (e.g. Deep Learning, K-Nearest Neighbour, Support Vector Machine etc.) to build models for forecasting traffic flow in short-term and in long-term. However, this field also study to predict traffic velocity, gap distance and traffic oscillation to improve the road traffic condition in smart cities around the world. This chapter presents recent studies on traffic flow prediction and some other works regarding the field of traffic data mining.

2.1 Traffic flow prediction

Predicting road traffic flow is important for better regulation of road traffic. A basic traffic flow prediction approach considers previous time-series traffic flow as input to generate a location prediction and trains machine learning methods. Different machine learning methods are being consistently used to improve prediction performance. The most recent ones will be discussed in this section.

2.1.1 Deep Learning / Neural Networks

Yisheng et. al. [11] presented a traffic flow forecast method based on deep learning approach with a stacked diabolo network or auto-encoder. This stacked auto-encoder was used to build a SAE (Stacked auto-encoder) model where, a logistic regression was applied on the top layer as standard predictor. Which helped the network to learn about its traffic flow features. The study collected three months traffic data of California using PeMS database. The last month's data was considered as test dataset. Their model performed significantly good in comparison with BP NN (Backpropagation NN, RW (Random Walk), SVM, and RBF NN (Radial basis function based NN model).Yanjie et. al. [12] extend the work with further evaluation of the auto-encoded SAE model at day and night time. The result show that the performance of the model is not similar for day and night time, indicating requirement for further improvement. Teng et. al. [13] came up with a boosting up scheme to eliminate the shortcomings of SAE. In this work, an algorithm using replication strategy was proposed to train multiple SAEs. The result indicated better forecasting performance.

Arief et. al. [14] studied correlation between parameters of weather conditions and traffic measurement and incorporated that insight into DBNs (Deep Belief Networks) to bring accuracy in prediction of traffic flow. A traffic dataset from the Bay area of San Francisco was analysed for correlation study. The study evaluated that DBNs predict better but takes more time in comparison with ARIMA and ANN. As a deep learning approach LSTM (Long short term memory networks) models can also be used for predicting road traffic. Like, Zheng et. al. [15] and Hongxin and Boon-Hee [16] both build LSTM models using deep learning approach to predict short-term traffic flow. The error rate were relatively low for both models. The capability of learning the architecture from non linear traffic data and the ability of capturing dependencies in long-term from time series data make LSTMs suitable for traffic forecast. The performance of deep learning model with linear vector autoregressive was compared by Nicholas and Vadim [17]. The study collected dataset from Chicago metropolitan area and applied several data pre-filtering techniques to compare the performance. DLM8L (Deep learning with median filtering processing lasso) performed the best among them.

ANNs (Artificial Neural Networks) uses human brain principles perceive information or pattern from unknown data sources. Many studies continuously applying Neural Networks (NNs) combined with other techniques for non-linear traffic flow prediction. Such as, Rui et. al. [18] used LSTM and GRU (Gated Recurrent Units) for training neural networks. LSTM and GRU both had forgetting units. Both neural networks outperformed ARIMA model in terms of predicting traffic flow of Alameda and Oakland, California. A Mallat wavelet analysis based back propagation NN was proposed by

Chai et. al. [19]. Their model produced better result for the traffic flow of Chongqing, Chengdu in comparison with moving average prediction model. Se-do et. al. [20] also trained a neural network applying back propagation with Levenberg Marquardt (LM) algorithm for traffic flow prediction. The study used data from Suwon city, South Korea. An innovative multifactor pattern recognition model (MPRM) was introduced to determine the weights of each variables. Which, later gets integrated with GMM (Gaussian mixed model) clustering and an ANN was applied on each cluster for the prediction.

2.1.2 Nearest Neighbour

The ability of K-NN (K nearest neighbor) algorithm often gets optimized for short-term traffic flow prediction. Meng et. al. [21] proposed a short-term traffic flow prediction model by improving the traditional KNN algorithm. The study have utilized pattern recognition 2 times when searching for matches in the case database. The case database was created with the self balanced binary tree technique called AVL (Adelson-Velsky and Landis). The efficiency of their model (AKNN-AVL) has been proved using dataset from North 3rd Ring Road, Beijing. Marcin et. al. [22] improved traditional KNN for prediction with data segmentation method that is capable of selecting useful input data. Usefulness of the data was validated through a similarity check and by ranking of the prediction error for traffic flow. Their model was evaluated with the traffic flow of Gliwice, Poland. Dawen et. al. [23] proposed an innovative Map Reduce based nearest neighbor framework and performed an empirical study on the road traffic flow of Beijing. Pietro et. al. [24] also tried to strengthen nearest neighbor regression with multivariate and a time awareness approach. Performance of their method was examined over the dataset collected from California Highway. A new hybrid multiple metrics based KNN called HMMKNN model was introduced by Haikun et. al. [25] for traffic forecasting. The model had the ability to extract intrinsic property from data. The accuracy of HMMKNN was better than the other traditional models. Jing and Shiliang [26] incorporated KNN algorithm with gaussian process dynamic model for efficient traffic flow prediction. Basically, KNN algorithm was introduced for the prediction of latent variables. The predictions generated by the 4th order gaussian process dynamic model and KNN were averaged for the forecast estimation.

2.1.3 Support Vector Regression

Support Vector Regression (SVR) and Support Vector Machine (SVM) use pretty much the same learning methods in terms of classifying data. With some modification SVM and SVR can perform better traffic flow prediction and they are highly applicable for better parameter selection. Xianyao et. al. [27] developed an algorithm using Multi kernel Support Vector Machine (MSVM) to predict road traffic flow in short-term. To optimize the parameter for MSVM the study also proposes an Adaptive particle swarm optimization algorithm. Mean squared error was used to calculate fitness value among the particles of the given dataset. The dataset was collected through roadside units from PeMS system. Five days data from a week (monday to Friday) were collected to examine the performance with comparison to other proposed SVM methods. Yanyan et. al. [28] prepared an urban traffic flow forecasting model by training SVR on some weighted predictors. The predictors were selected using multivariate adaptive regression splines which is a non parametric regression approach. The model was experimented over the traffic flow of Shanghai, China and the model had a good accuracy. Xiaobo et. al. [29] developed an algorithm SHGA (sparse hybrid genetic algorithm) to handle variable selection problem of LSSVR (least squared SVR). A hybrid encoding scheme was prepared with a sparsity constraint to limit number of variables engaged in LSSVR. This approach helped to gain better performance in predicting traffic flow providing fewer spatiotemporal variables to deal with.

Jinyoung et. al. [30] presented a real time road traffic forecasting model using Bayesian classifier and SVR. The flow of traffic and their relation was modeled using 3D Markov random field (MRF). The study defined some cliques based on those relations which get estimated utilizing multiple linear regression and SVR. The study colled traffic dataset from Gyeongbu expressway, South Korea and the traffic flow was predicted based on the traffic speed for the next time-stamp. Afterwards, E Y Kim [31] further evaluates the model's generalisation efficiency by comparing it with some existing works.

2.1.4 Other approaches

Many different approaches have been proposed in recent times to predict traffic congestion. Many proposes new techniques to process data to help machine learning algorithms to predict better.

Bayesian Networks:

Bayesian networks are very promising to undearstand relationship between scattered variables. Shiliang et. al. [32] structured a Bayesian network based on the links of adjacent road from a road traffic network. Then, the Bayesian network was utilized to predict values of the target nodes. Which, then get expanded by GMM (Gaussian mixed model) and PCA (Principal component analysis) for assessing the probability joint. The performance of their expanded Bayesian network was better than AR (Autoregressive) method over an incomplete data of Beijing City. Xun et. al. [33] also used Bayesian algorithm to build a hybrid model to improve performance of SARIMA (Seasonal autoregressive integrated moving average) model for traffic flow prediction. With a real time dataset from New York their proposed model outperformed some existing schemes. Zheng et. al. [34] combined LCG (Linear conditional Gaussian) and Bayesian network to forecast traffic. Both of the models accept continuous and discrete variables. This allow categorical variables to consider for forecasting. Comparing with other approaches resulted good possibility of their model for a simulated traffic network.

Autoregression:

Dong-wei et. al. [35] combined ARIMA (autoregressive integrated moving average) model with Kalman filter to build a traffic forecasting algorithm. The algorithm is capable to detect optimal parameters from historical traffic flow. Traffic dataset was collected from Beijing, China and the algorithm performed better over ARIMA and Kalman filter models.Hubert and Jerome [36] also proposed a private traffic estimation algorithm which was based of Kalman filtering. Afshin et. al. [37] developed an adaptive autoregressive model based traffic forecasting algorithm. The study used dynamic traffic simulator where an OD (Origin-destination) matrix was used for data completion. Link to like driving ratio was used by the algorithms to perform prediction over the traffic flow of San Francisco, California. Another traffic flow forecasting model using Topology regularized vector autoregression named TRU-VAR was proposed by

Florin et. al. [38]. The study used traffic datasets of Melbourne, Australia and California freeway. The model outperformed ARIMA and CMSD (Contextual mean seasonal difference) baseline model for different datasets.

Logistic Regression:

Suguna and Neetha [39] came up with an algorithm that can predict traffic congestion by training machine learning algorithms. This was based on the grouping data generated after examining the average speed of the road. Paper showed that logistic regression performed the best for their algorithm. Dick et. al. [40] estimated road traffic volume using Linear and Logistic regression models for Wyoming, US. The inconsistency of error variance was taken care of by the log transformation of response variable. Both models were cost efficient and had a high accuracy.

Ensemble methods:

Yalda et. al. [41] combines HW (Hall White) Model and EV (Extended Vasicek) Model together for better performance in traffic flow forecasting. The work used datasets from Tehrans highway and an open access PeMS database. The HW model considers both short-term and long-term data to generate a baseline prediction. The prediction with HW model did not perform good for an occasional day. Therefore, they introduced a new step by applying EV model to predict the residuals. EV model is a stochastical process which uses a drift function and variance to perform mean characterization. The performance of their proposed model was better than HW model and some other machine learning methods.

Zhiyuan et. al. [42] collected four years data from Phoenix, US to build an adaptive traffic forecasting model using ELM (Extreme learning machine). The initial prediction was of the model was good for large data but not reliable for small dataset. To overcome the problem a new scheme using OSELM (Online Sequential ELM) was proposed into the paper to structure a new ELM network from a small dataset. This helped the model to produce better accuracy. Ridha et. al. [43] proposed a short-term traffic flow prediction framework that used data fusion to manage data heterogeneity. The study incorporated weather and traffic data and also used social media data (such as, tweets) to propose a DBN (Deep belief networks) model. The model was experimented using dataset of San Francisco from PeMS database.

Athanasios et. al. [44] proposed a model for short time traffic flow prediction considering both normal and abnormal road traffic events. They identified traffic pattern occurred on particular VDS (Vehicle detection stations) from State of California. Then, using that knowledge the data was transformed into 3D feature vectors and a density-based data clustering algorithm was applied to find clusters. Afterwards, they applied their technique over ARIMA, SVR and kNN which outperformed the existing prediction models. Asif and Ming [45] also proposed genetic algorithm (GA) based locally weighted regression models (GA-LWR) and compared the performance with another GA based ANN. GA algorithm was used to optimize input dataset collected from Beijing, China. Their model outperformed GA based ANN models.

2.2 Traffic flow optimization

The main objective of ITS is to control the flow of vehicles on roads for better transportation system. Therefore, many scholars propose traffic control policies to regulate traffic congestion.

A traffic control policy for an assumed large scale vehicular network was proposed by Alfred et. al. [46]. The network outflow was designed using the NFD (Network Fundamental Diagram) approach. And the network inflow followed the admission control policy. The goal was to maximize the network throughput maintaining the balance between the external network and the internal network. The work used three different controllers (PI controller, MPC controller and relaxed controller) over a test model based on Stockholm, Sweden. The experimental result indicated that PI controller performs better until the external network saturate. Relaxed controller performs better when constraints are conflicting in the internal network. This result can improve the flow of the vehicles by enhancing network throughput for road traffic. Sasan et. al. [47] analyzed the needs of big data architecture for controlling road traffic and propose a new architecture. The architecture divides the publishers and the subscribers from ITS actors (camera, Bluetooth etc) by implementing kafka. Kafka basically is a cluster also known as Apache kafka helps to manage stored information or messages between producers and customers. Kafka helped to handle the vast bandwidth of incoming data stream. The architecture also stores information in a Hadoop file system for further analysis. They have performed a case study on a freeway of north of Munich, Germany.

The study used SUMO (Simulation of urban mobility) for better representation of the experiment. Using python programming they provided control logic for opening and closing of a hard shoulder (emergency stand for cars at a breakdown on the road). Ricardo et. al. [48] and Samuel et. al. [49] both study presented traffic control approaches based on traffic flow prediction models with the collected datasets from different states of United States.

Mofan et. al. [50] aimed to use recurrent neural network (RNN) to predict road traffic oscillation. They investigated that insufficient data and inaccurate measurements cause the neural network to perform bad to predict traffic oscillation. However, with sufficient field data the model produced highly accurate results. The dataset was collected from Emeryville, California using Next Generation Simulation (NGSIM). Data preprocessing was performed using moving average smoothing. Utilizing their investigation a new recurrent neural network architecture (RNNa model) was proposed at the trajectory level to improve neural networks performance. The study tests their models performance with traditional car flowing model IDM (Intelligent Driver Model) and their model outperformed the IDM model in predicting the behavior of different type of drivers.

2.3 Traffic Information Extraction

Improving road traffic condition requires proper understanding of road traffic. Data scientists consistently applying different techniques to extract information from the roads, such as, analyzing drivers choice behavior, travel behavior, predicting vehicles speed on the road and so on.

2.3.1 Road Traffic Condition

Nik et. al. [51] tried to study the overall outflow of three major highways of Malaysia. The work used different traffic stream models over the provided data and understood the behaviour of one parameter to another in terms of speed, density and traffic flow. Trans logarithmic function was used to analyze data for understanding the congestion level of each route. Ibai et. al. [52] characterizes the road traffic flow of Madrid, Spain using a density based clustering algorithm. The study tried to find commonalities among the instances by making clusters. Then, it used the clustering technique to improve traffic prediction performance. Although, the study concluded with room for improvements but, the approach was capable to identify patterns for an urban road scenario. Chao et. al. [53] proposed an algorithm to observe and select parameters from traffic flow models. The main approach of their work was assimilating Eulerian (road sensors data) and Lagrangian (GPS data) observation models data and interpret them. Their model was able to estimate traffic state for shorter range traffic flow. But, further improvement was required for wider range traffic flow.

Alireza et. al. [54] put together Percolation theory and linear string stability analysis to understand the effect of information availability for better traffic flow. Continuum Percolation models often get used in wireless communication which is connected to discrete percolation model to find critical density. The study investigated this approach to find critical density among the vehicles on the road by using two lemmas. The average distance among k-components was calculated using these lemmas considering a Poisson distribution for road vehicles. Finally study establishes the significance of using Percolation theory in vehicular ad-hoc network. The study also identified the relationship the critical density and communication range.

2.3.2 Behavior Analysis for Road Traffic

Mauro and Mario [55] proposed a possibility theory based model to evaluate the drivers behavior of choosing en-routes. Possibility theory is related with fuzzy sets theory which has a mathematical tool to examine the possibility of occurring any incident. Their proposed model uses the available information from ATIS (Advanced traveler information system) and it updates drivers perception considering a time-dependant possibility distribution when finds a new information. They evaluated their model using a driving simulator for computer (UC-win/Road) at Polytechnic University of Bari. The experiment took place for 6 months. RMSE for the proposed model was found to be respectively 5.36% and 10.0% for queue message and accident message which proved the model to be effective. Cynthia et. al. [56] introduces the similarities between the travel behavior analysis with the big data analysis. This work shwed researchers from travel behavior analysis and researchers from big data analysis have some cross-disiplinary ideas. To achive their goal both areas develop models with similar idology, analyze behavioral factors with movement patterns. Collaboration of these two fields is important for the efficient analysis of road travel behavior.

2.3.3 Vehicular speed prediction

Zheyuan et. al. [57] introduced a vehicle speed prediction algorithm based on big data and deep learning. Data from different fields were considered such as, route information, driver behavior, traffic and weather data. The study used supervised learning by training an adaptive Nero fuzzy network with 5 layers. the algorithm was tested by a driver where, data was collected using GPS device. Jinjun et. al. [58] also trained a fuzzy neural network (FNN) which they have used to predict traffic speed of Beijing city using one trigonometric regression function. Input samples in the clusters were partitioned using K means. Then, Gaussian fuzzy membership function was utilized to measure the cluster centers. A weighted RLS (Recursive least squares) filter algorithm used to optimize parameters of fuzzy rules. Afterwards, the trigonometric regression fitted the periodic patterns to improve performance of FNN. Changhee et. al. [59] proposed a traffic speed prediction model for main roads of Seoul, South Korea. A CNN (Convolutional Neural Network) model and two MLP (Multilayer perceptron) models were build to estimate the traffic speed. CNN model was created with 5 input layers and it outperformed the other two models based on their experimental result.

Baozhen et. al. [60] developed a short term traffic speed prediction model along with a single step prediction model using SVM model. The model calculates the arrival time from one route to the next and also estimates speed on the next road link. Thus, computes the destination arrival time. Another traffic time prediction model was proposed by Matej et. al. [61]. Their methodology and system architecture was based on a multivariate hybrid method. The study used PPCA (Probabilistic principal component analysis) to identify correlation patterns among the links of Stockholm, Texas.

2.4 Summary

Different traffic flow prediction models are required for different types of environment. One approach producing good result at any particular area might not produce similar result in another area. Therefore, it is not possible to conclude any universal model that would fit everywhere in the world. So, this research area of forecasting traffic should be explored more. And, we can see, most of the studies are based on short-term traffic flow prediction which, generally indicates 5 to 30 minutes. Therefore, more models should be build focusing on long-term traffic flow prediction.

Chapter 3

Dataset

Accurate collection of road traffic data is inevitable for better ITS to help make smart cities. Since the last decade, technologies give access to real world traffic data through different methods and applications. This chapter will discuss about the traffic dataset collected for this study and the processing techniques applied for real world traffic flow prediction.

3.1 Data collection

The real-time road traffic data was collected from Porto city, Portugal. Porto is the second largest city of Portugal with a population of 2.4 million. The city has several one way roads and the roads are fairly congested. However, the loop sensors to collect the road traffic data were mostly placed in VCI (Via de Cintura Interna), Porto. It is a ring shaped motorway signposted as IC23. VCI is 21 km long and it encloses municipalities of Porto and Vila Nova de Gaia together. Vila Nova de Gaia is known as the most affluent suburb with a population of more than three hundred thousand. Therefore, everyday a large number of vehicles circulate this motorway. Loop sensors, also known as inductive loop detectors are very widely used for collecting information of road traffic. And the installation process is fairly easy. There were 22 sensors to sense vehicular information from the road. Each row of the dataset is calculated for every five minutes. After sensing the vehicular data from the roads through loop sensors the information get stored in a PostgreSQL database. This is a strong, reliable and SQL standard open-source database management system. The data is accessible for research purposes. Our initially collected traffic data had three years of road traffic information, from the year 2013 to 2015. Table 3.1 describes the columns or features of the traffic dataset. The primary objective of this study was to forecast road traffic flow based on the historical traffic data. The month considered in this paper was April, 2014 which had 23 attributes with 10,39,077 instances.

Table 3.1: Feature description of the traffic dataset.

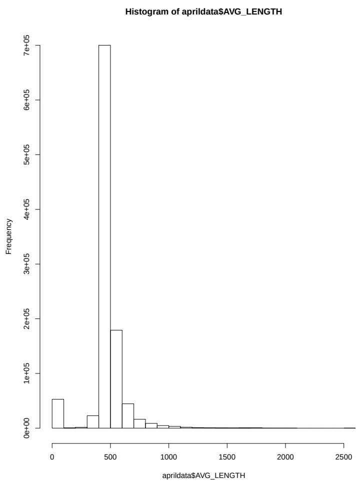
Name	Description
EQUIPMENTID	Sensor ID
LANE_NR	Lane number
LANE_DIRECTION	Lane direction
TOTAL_VOLUME	Number of vehicles
AVG_SPEED_ARITHMETIC	Arithmetic average speed of vehicles
AVG_SPEED_HARMONIC	Harmonic average speed of vehicles
AVG_LENGTH	Average length of the vehicles
AVG_SPACING	Average spacing between vehicles
OCCUPANCY	Occupancy
INVERTED_CIRCULATION	No. of vehicles pass through with opposite direction.
LIGHT_VEHICLE_RATE	The % is the ratio of light vehicles
VOLUME_CLASSE_A	No. of vehicles of class A
VOLUME_CLASSE_B	No. of vehicles of class B
VOLUME_CLASSE_C	No. of vehicles of class C
VOLUME_CLASSE_D	No. of vehicles of class D
VOLUME_CLASSE_0	No. of unidentify vehicles
AXLE_CLASS_VOLUMES	No. of vehicles per vehicle axis number.
AGGREGATE_BY_LANEID	Is a unique ID fusing two IDs, the Lane ID and the Lane Direction.
AGG_PERIOD_LEN_MINS	Is the time between each measurement the sensors made. It is always 5 minutes.
AGG_PERIOD_START	Time interval
NR_LANES	No. of lanes in the street of the event.
AGGREGATE_BY_LANE_BUNDLEID	Similar to AGGREGATE_BY_LANEID
AGG_ID	Street ID

3.2 Data Visualization

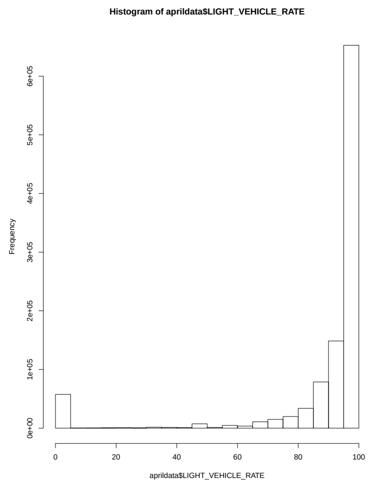
The study has used two most popular programming languages for statistical data computing, "R Programming" and "Python" to draw different graphs based on the attributes, such as, histogram, density graph and pie chart to understand the traffic data.

3.2.1 Frequency Analysis of Attributes

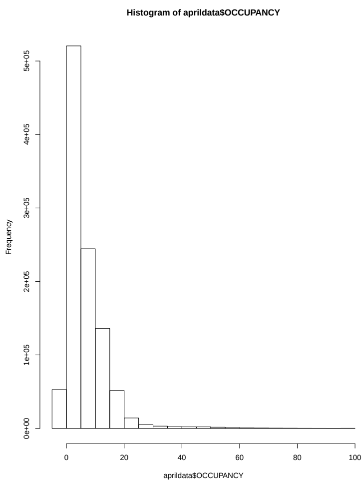
Histogram of some of the attributes of the traffic data are shown below in figure 3.1:



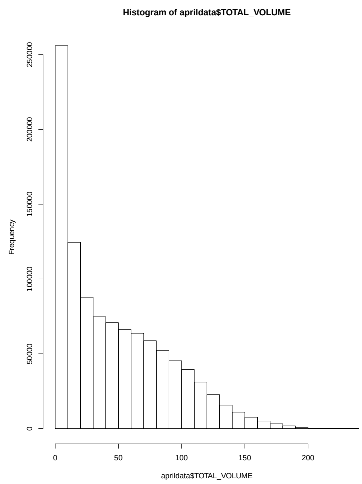
(a) Histogram for Average Length of the cars.



(b) Histogram for Light Vehicle Rate.



(c) Histogram for Occupancy.

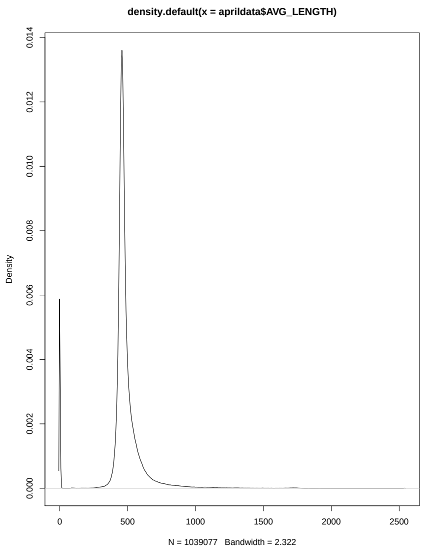


(d) Histogram for Total Volume.

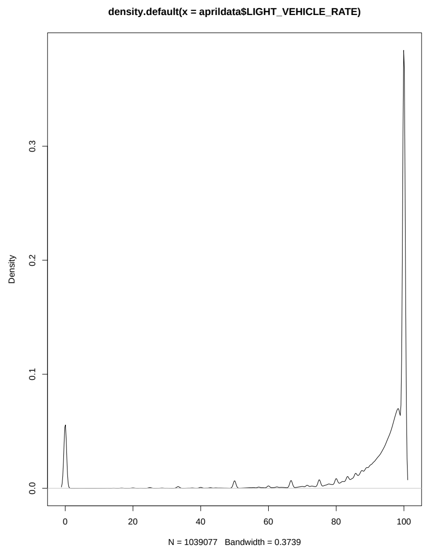
Figure 3.1: Histogram for some of the attributes.

Density graph of some of the attributes of the traffic data are shown below in figure

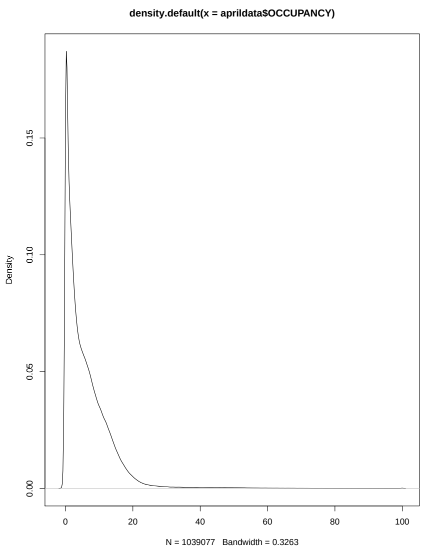
3.2:



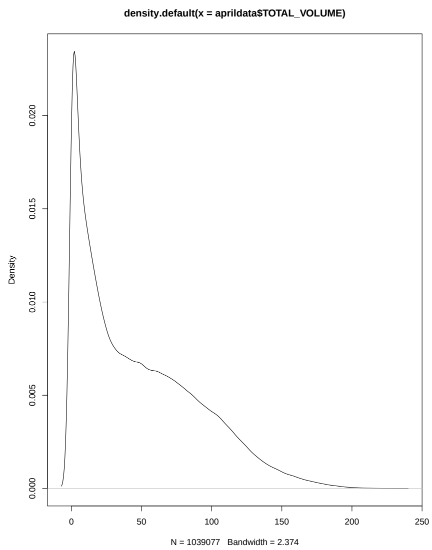
(a) Density graph for Average Length of the cars.



(b) Density graph for Light Vehicle Rate.



(c) Density graph for Occupancy.



(d) Density graph for Total Volume.

Figure 3.2: Density graph for some of the attributes.

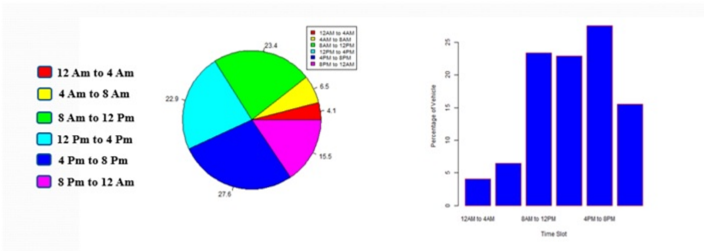


Figure 3.3: Pie Chart (left) and Bar Chart (right) describing percentage of traffic in different time slots of the month April, 2014.

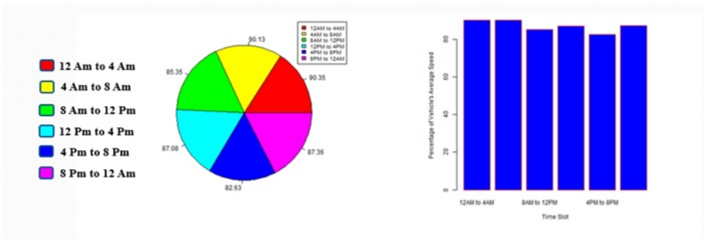


Figure 3.4: Pie Chart (left) and Bar Chart (right) describing average speed of the vehicles in different time slots of the month April, 2014.

3.2.2 Road Traffic Flow Visualization

We have plotted different pie charts and bar charts to perceive information about the traffic condition of the city. These diagrams are shown in figure 3.3 and figure 3.4.

Figure 3.5 shows a bar chart of Total Volume with other Volume Classes for April, 2014. And Figure 3.6 Describes total number of vehicles pass through each sensor of the city for the whole month.

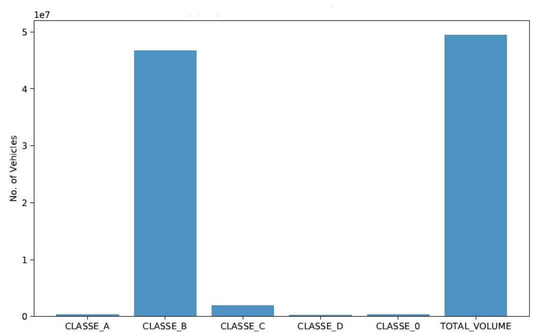


Figure 3.5: Number of vehicles along with car classes for the Month April, 2014.

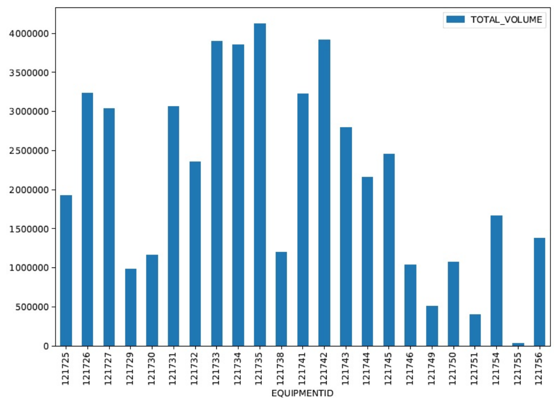


Figure 3.6: Equipment Id vs Total Volume of vehicles of the Month.

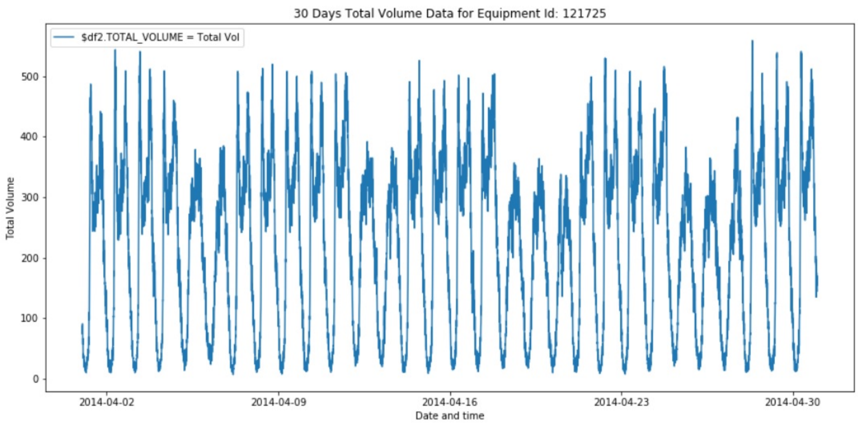


Figure 3.7: Month April, 2014.

Additionally, The overall traffic flow of the month for a particular road with equipment id: 121725 is showed in figure 3.7:

3.3 Data Preprocessing

Data Preprocessing is one of the major steps of data mining to extract knowledge. Our goal was to prepare the collected dataset to apply regression models for predicting the road traffic flow. This study focused on the month April, 2014 and predict the fourth week's traffic flow based on previous three weeks of traffic flow of VCI, Porto city. As mentioned earlier, the dataset had more than one million instances and the randomness in the real world traffic data made it difficult to visualize it properly. Therefore, we followed some fundamental steps of data preprocessing to understand and prepare our collected traffic dataset for prediction.

3.3.1 Data Cleaning

Data cleaning or data cleansing is indispensable for accurate and reliable prediction of data. Our data cleansing process was attained using pandas library from Python programming. The library contains loads of data manipulation techniques for real world data analysis. The traffic data of VCI, Porto city had no missing values. We serialized the dataset utilizing the data sorting technique to clear the randomness and smoothing the data.

3.3.2 Data Integration

Data integration depends on the principles of how one want to sync the data. We wanted to gain knowledge from the collected traffic dataset and instruct different regression models to perform the predictions. The collected dataset had already been integrated with the vehicular information sensed by 22 different sensors on the roads. Each sensor has an unique equipment id. Figure 3.5 shows a 3D graph representing the EQUIPMENTID vs the LANE_NR vs the TOTAL_VOLUME for April, 2014. This figure reveals a valuable information for the study, that is, the number of lanes differ for each of the equipment ids placed on different roads of the city.

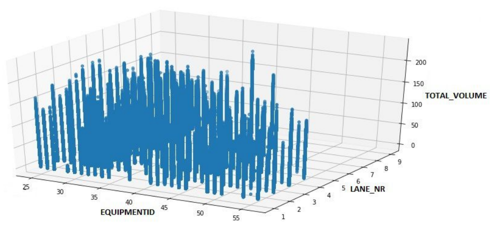


Figure 3.8: EQUIPMENTID vs LANE_NR vs TOTAL_VOLUME - April, 2014, Porto city.

3.3.3 Data Transformation

The study wanted to predict the traffic flow on different roads of the city separately. We mapped the features of our dataset to plan the aggregation approach. As each road has multiple lanes, therefore, we summed up the vehicular flow of lanes for each of the roads. We also summed up the total volumes of the traffic flow for each day of the month to make it easier for the regression models to forecast the flow of future road traffic.

3.3.4 Data Reduction

Real world data contains loads of redundant information. Data Reduction is a must done step to reduce the complexity of analyzing real world data without hampering the target experiment. We analyzed the features of our traffic dataset to find dispensable attributes. Table 3.2 reveals some non-essential features of the traffic dataset, like, INVERTED_CIRCULATION, AGG_PERIOD_LANE_MINS. We also analyzed the correlation coefficient between two attributes. In Table 3.3 we try to understand the correlation each attributes has with Total Volume. In this study, we focus on forecasting the total volume of the traffic flow on each road. Therefore, we excluded other volume classes from our consideration as well.

3.3.5 Data Discretization

In machine learning or statistics, data discretization is the method of converting continuous data or features to a number of discrete or nominal counterparts. This process reduces noise in the data also minimizes size and smooths the dataset. Decretization methods can be performed manually with distribution analysis or automatically using binning, regression analysis, cluster analysis etc. Although, it is a useful process but in some cases reforming the data into chunk of bins can distort real data. Fitting parametric models using standard statistical principal, such as, maximum likelihood can be applied in those cases. However, this study avoid any manipulation of traffic data to test the accuracy of the regression models on traffic flow prediction.

Table 3.2: Feature Statistics

Attribute Name	Min Value	Max Value	Mean Value	STD Value
EQUIPMENTID	121725	121756	121737.683	8.639
LANE_NR	1	9	3.621	2.036
TOTAL_VOLUME	0	233	47.62	42.125
AVG_SPEED_ARITHMETIC	0	255	82.615	25.553
AVG_SPEED_HERMONIC	0	255	81.098	25.491
AVG_LENGTH	0	2540	470.569	151.101
AVG_SPACING	0	4642.06	348.129	454.89
OCCUPANCY	-1	100	6.221	7.459
INVERTED_CIRCULATION	0	0	0	0
LIGHT_VIHICLE_RATE	0	100	89.104	23.521
VOLUME_CLASSE_A	0	62	0.324	0.832
VOLUME_CLASSE_B	0	233	44.905	40.105
VOLUME_CLASSE_C	0	34	1.812	3.134
VOLUME_CLASSE_D	0	13	0.238	0.619
VOLUME_CLASSE_O	0	110	0.341	1.6
AGGREGATE_BY_LANEID	14903478	16000350	15460586.469	312787.274
AGG_PERIOD_LANE_MINS	5	5	5	0
AGG_ID	94373	1589493	1005673.608	359072.851

3.4 Feature Selection

Feature reduction process has already been mentioned earlier in this chapter. We concluded with four major features with one class feature to train the regression models. The features are: i. EQUIPMENTID, ii. AGG_PERIOD_LANE_MINS, iii. LANE_NR, and iv. TOTAL_VOLUME. The study used these features to train regression models that fit a pattern based on the historical traffic flow and use the pattern to forecast the future traffic.

3.5 Summary

This section briefly described the traffic data that we have collected for our study. This was collected by 22 loop sensors, that constantly sensed vehicular information from the roads of VCI motorway, city of Porto, Portugal. We have considered the roads data of the month April, 2014 for this experiment. We have also discussed about the data visualization and data preprocessing steps concisely in this section.

Table 3.3: Correlation Coefficient of Attributes with Total Volume

Name of Attribute	Correlation with TOTAL VOLUME
EQUIPMENTID	-0.0771
LANE NR	0.01148
TOTAL VOLUME	1
AGG PERIOD START	-0.01551
AVG SPEED ARITHMETIC	0.0121
AVG SPEED HARMONIC	0.011005
AVG LENGTH	0.138
AVG SPACING	-0.57067
OCCUPANCY	0.79149
LIGHT VEHICLE RATE	0.282436
VOLUME CLASS A	0.345309
VOLUME CLASS B	0.996278
VOLUME CLASS C	0.47663
VOLUME CLASS D	0.281426
VOLUME CLASS 0	0.133457

Chapter 4

Regression Analysis

Regression analysis is a well known method for understanding the dependency among variables. This variables can indicate to features or attributes. The dependency or relationship among attributes is estimated by a set of statistical process. The mathematical relation within the variables is then used by the regression models for estimating value for an independent data point.

4.1 Linear Regression

Linear regression is a supervised learning algorithm which uses linear approach for prediction problem [62]. Linear regression try to fit a line that explains the relationship between the independent variable and the dependent variable. If dependent variable increases with the increment of the independent variable then it is refereed as positively related. And if the relationship is opposite then it is called negatively related. The goal of a linear regression line is to have minimum error for all the data points.

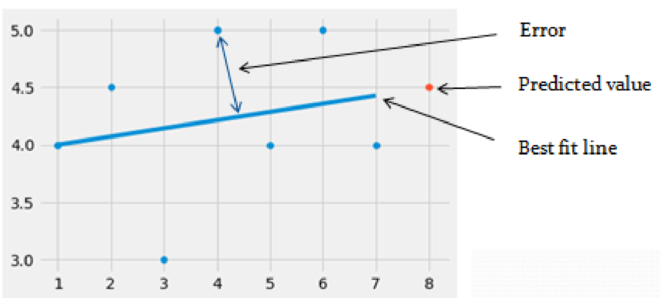


Figure 4.1: Linear Regression - Prediction with best fit line.

Calculation of Linear Regression:

Simple linear regression follows the traditional equation of a line,

$$y = mx + c \tag{4.1}$$

If the mean of x values is $\bar{x}$, and the mean of y values is $\bar{y}$, then the slope m is calculated with the formula given below,

$$m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

And the y-intercept c is calculated by subtracting the mean of y values with the mean of x values multiplied with m:

$$c = \bar{y} - m\bar{x}$$

Therefore, equation (4.1) can be used to predict any data point of x. Linear regression is simple to implement and it is used to predict numeric values.

4.2 SMO Regression

Sequential minimal optimization or SMO regression is used to train SVM (support vector regression). In 1998, the idea of breaking the large quadratic programming (QP) problem for training SVM was first proposed [63]. Afterwards, this concept was taken into account for solving regression problem as well [64]. With time many researchers have suggested approaches to improve the performance of the SMO algorithm for SVM classifier [65]. SMO algorithm optimizes two variables at a time on each iteration which finds the optimum analytically without using numerical optimization techniques. So, a major part of the algorithm works on choosing the two variables for each iteration. This made easy and fast to solve SVM or QP optimization problems.

Quadratic Programming Problem:

A general quadratic programming problem has a quadratic objective function with linear constraints. A basic form is shown below:

$$\begin{array}{ll} \underset{X}{\text{minimize}} & \frac{1}{2} * X_i^T * C * X_i + g^T * X_i \\ \text{subject to} & 0 \leq X_i \leq H,\ i = 1, \ldots, m. \end{array}$$

Where, X = $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ is a vector of the decision variables, H = $\begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix}$ contains the constant coefficients that would multiply the squared terms and the x1 times x2 terms. So, the H needs to get multiplied two time by the column vector X. The linear terms in the model is defined by g = $\begin{bmatrix} g_1 \\ g_2 \end{bmatrix}$, the coefficients are g1 and g2 in the vector. 0 is defining the lower bound and H is the higher or upper bound in the constraint. The linear constraints also get defined in the model.

Calculation of SMO Regression:

SMO regression can be considered as an improved algorithm for SVM. Dual SVM optimizes all the n variables $(\alpha_1, \alpha_2, \alpha_3, \ldots \alpha_n)$ with applying the kernel trick. Here, $y_i = \pm 1$, which refers to the class that each x_i is assigned.

Dual SVM Optimization,

$$\begin{aligned} \underset{\alpha}{\text{minimum}} \quad & \frac{1}{2} \sum_i \sum_i \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_i \alpha_i \\ \text{subject to} \quad & \sum_i \alpha_i y_i = 0, \text{ and } \alpha_i \; \epsilon \; [0, C] \end{aligned}$$

SMO algorithm instead of optimizing all the n variables only optimizes two variables (α_a and α_b) at each iteration. K_{ij} here representing the kernel function,

$$\begin{aligned} \underset{\alpha_a, \alpha_b}{\text{minimum}} \quad & \frac{1}{2} \sum_i \sum_i \alpha_i \alpha_j y_i y_j K_{ij} - \sum_i \alpha_i \\ \text{subject to} \quad & \sum_i \alpha_i y_i = 0, \text{ and } \alpha_i \; \epsilon \; [0, C] \end{aligned}$$

SMO Algorithm basically works with two steps:
1. Find the most promising pair (α_a and α_b).
2. Optimize α_a and α_b keeping other α's fixed.
Until convergence the algorithm keeps repeating these steps in each iteration.

4.3 Multilayer Perceptron

Multilayer Perceptron also known as MLP is the idea of linking multiple perceptrons together to generate numerous outputs. Perceptron takes inputs into the processor along with weights and a bias to create a weighted sum. The weighted sum gets passed through an activation function which generate the output. A single perceptron can only solve the linearly seperable problem [66]. MLP has the ability to solve real world problem that are not linearly separable [67]. The activation function mentioned above is a sigmoid function denoted by σ and the formula is described below:

$$f(x) = \frac{1}{1+e^{-x}}$$
$$\text{Or, } f(x) = \frac{e^x}{e^x+1}$$

For a typical multilayer perceptron x is the weighted sum calculated from the bias and the connection of inputs.

Calculation of Multilayer Perceptron:

The inputs $(X_1, X_2 \dots X_n)$ can be considered as a 1-dimensional vector. And the weighted matrix W = $(W_{11}, W_{12} \dots W_{Nn})$ gets multiplied with input and summed up with the bias (B).

$$\begin{bmatrix} W_{11} & W_{12} & \dots W_{1n} \\ W_{21} & W_{22} & \dots W_{2n} \\ \vdots & & \\ W_{N1} & W_{N2} & \dots W_{Nn} \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_n \end{bmatrix}$$

This equation can be expressed as,

$$(W_{ij} * I + B) \tag{4.2}$$

Where, I is the input vector and $i = 1, 2, \dots n,$ and $j = 1, 2, \dots n.$

The activation function or sigmoid function σ is applied to over the weighted sum (4.2). This produces the output which in the case of multiple layers processes some hidden outputs (H).

$$\text{Hidden outputs, } H = \sigma(W_{ij} * I + B) \tag{4.3}$$

These hidden outputs from the equation (4.3) then processed through neurons or perceptrons to execute the final output. A different bias along with the sigmoid function executes the final output.

$$\text{Final output, } O = \sigma(W_{ij} * H + B) \tag{4.4}$$

4.4 M5P Regression Tree

M5P model tree uses M5 algorithm to create base routine for the regression tree. A single linear function cannot always give better prediction. In some cases multiple functions are required to generate prediction. Regression tree can model the non-linear behavior of data by making multiple patterns to generate a proper prediction. Regression trees generate values at their final leaf nodes. However, in 1992, a new approach was proposed by Quinlan of having linear models at the leaf nodes of regression trees instead of values using M5 [68]. Divide-and-conquer technique was used to make tree models. And linear models were created using standard regression technique after testing the error rates for attributes using standard deviation reduction (SDR). The detail calculation of the process is mentioned in this section. Moreover, an improved version of the model M5' has been proposed later on for handling missing features data [69]. The models generated by M5 are strong and relatively comprehensible.

Calculation of M5 Based Regression Tree:

If, T is the set of assumed training examples, and $T_i = T_1, T_2, T_3 \ldots T_n$ are the subset of divided nodes or branches of selected values, then, the error reduction is calculated using divergence metric called SDR. The formula is describe below:

$$\text{Standard deviation reduction (SDR)} = SD(T) - \sum_{i=1}^{n} \frac{|T_i|}{|T|} * SD(T_i) \tag{4.5}$$

In equation (4.5) standard deviation (SD) is used for calculation the error reduction. M5 algorithm then selects the attribute for which the calculation suggests maximum error reduction. Next, the algorithm does pruning the tree by examining each non leaf nodes. Based on the error rate the algorithm either provides a linear model to the leaf or makes it a sub-tree. Each linear models are simplified with parameter estimation to reduce error. Finally, a smoothing process is applied to improve the models performance.

$$\text{Prediction Smoothing, P'} = \frac{n_i * P(T_i) + K * MP(T)}{n_i + K} \tag{4.6}$$

Here, Ti are the braches of the subtree T. And, n_i refers to no. of training examples. P(T_i) is the predicted value passed to the node and MP(T) is the model predicted value at T. A smoothing constant is given with a default value of 15, denoted as K.

4.5 Random Forest

Random forest generates a lot of decision trees to make prediction for classification and regression problems. It uses ensemble learning technique, that is training several models or learners and combining their decisions to make an estimation. The first algorithm of creating forest with multiple trees was proposed by Tin in 1995 [70]. The algorithm enhanced generalization accuracy for decision tree based classifier and the performance of the algorithm was very promising for recognizing hand written digits. An improved random forest algorithm was developed by Leo to enhance its performance [71]. The study suggests good performance of the algorithm in classification and regression with random features and inputs. Random forest applies bagging technique for training or learning models. This technique is also known as bootstrap aggregating.

Calculation of Random Forest:

Feature bagging is an ensemble learning method that is used by random forest algorithm. Given an example or training set, $S = s_1, s_2, s_3, \dots s_n$ and responses, $R = r_1, r_2, r_3 \dots r_n$. The algorithm randomly selects samples with replacement by repeating bagging (T times).
For, $t = 1, 2 \dots T$:
- Select S_t, R_t samples from S and R.
- Train the regression tree, f_t over S_t, R_t.

After the learning process predictions can be performed based on the regression trees on s'.

$$\hat{f} = \frac{1}{T} \sum_{t=1}^{T} f_t(s') \tag{4.7}$$

And, for classification the algorithm considers the majority votes based on the generated trees. Additionally, a standard deviation (SD) based on the predictions from generated trees can be calculated to learn about the prediction uncertainty.

$$SD = \sqrt{\frac{\sum_{t=1}^{T} (f_t(s') - \hat{f})^2}{T - 1}}$$

T is an independent parameter representing the number of trees which can be determined through cross-validation. After fitting a number of trees the training and testing error starts to saturate. Random forest can also be used for unsupervised learning by generating a predictor that can distinguish between observed data from synthetic data [72]. The ability of handling mixed attribute types and implicit variable selection make the algorithm suitable for large amount of data.

4.6 Proposed Ensemble Model

In this study, we propose an ensemble model by combining three regression algorithms that we have found most promising for predicting the traffic flow of Porto city. These regression algorithms are: i. SMO algorithm, ii. M5 algorithm and iii. Random Forest. Essentially, the model calculates the average predicted value of these regression models and make estimation for future traffic of the city. Figure 4.2 explains the overall method for the proposed model.

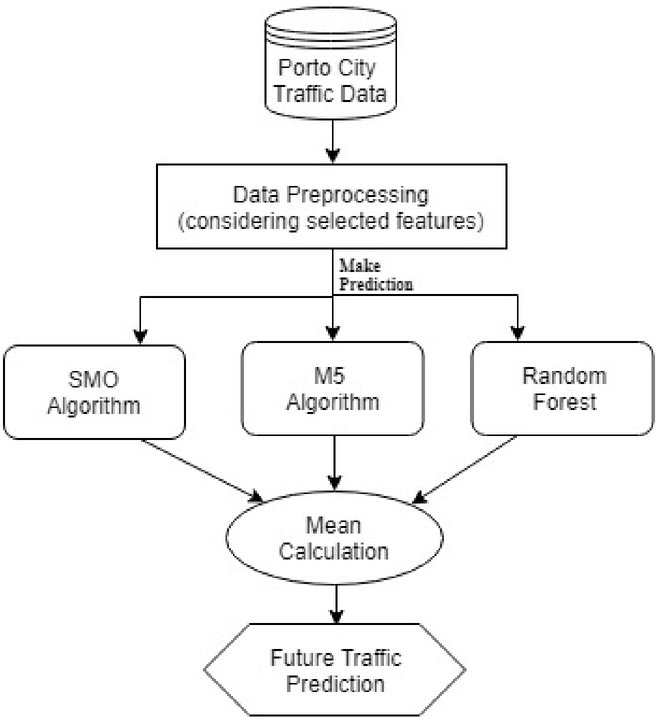


Figure 4.2: Proposed model based on Ensemble Learning Technique.

Chapter 5

Results

This study investigate the eligibility and capability of regression model on long term traffic flow prediction. Regression models that were considered are: i. Linear Regression, ii. SMO regression, iii. Multilayer perceptron, iv. M5P based regression tree model, v. Random forest and vi. Proposed Model.

5.1 Experimental Setup

This study applied different machine learning algorithms using Waikato Environment for Knowledge Analysis (WEKA) [73]. This software is written in java and it is widely popular for data mining tasks. The software is open source and it also gives the ability to call machine learning algorithms through java code using weka library. An additional package called 'timeseriesForecasting' was installed into the workbench to generate an environment for time-series prediction. This package transforms the original data into a large number of attributes to apply standard propositional learning algorithms.

The real time traffic dataset was already prepared to forecast according to the package requirement. We input first three week's traffic volume of the month April, 2014 to train the regression models. This study use these models to forecast the fourth week's traffic flow of the city. Therefore, the no. of units to forecast was set to 7. We also set the models to forecast on a daily basis. The linear model that was used for prediction was associated with M5 method for selecting attributes along with a rigid parameter set to default $1.0e^{-}8$. For SMO regression, we set the filter type as normalise training data for better data integrity. Other parameters were set as the default setup

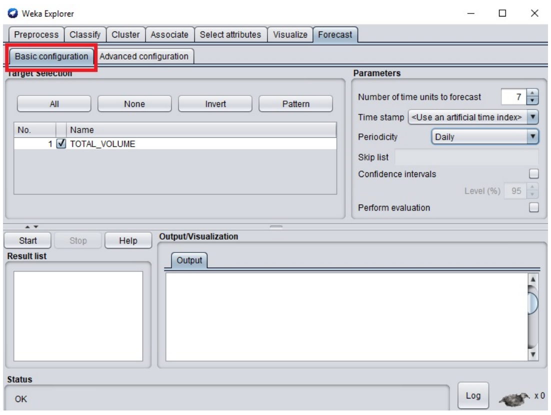


Figure 5.1: Basic configuration panel.

of the workbench. The lag variables were generated for better traffic flow prediction by the time series package itself.

5.2 Timeseries Forecasting Package

As mentioned earlier, timeseriesForecasting is an additional package of WEKA software that is capable of building time series forecasting models using machine learning algorithms and evaluate their performance. The basic and advanced configuration of the package have been briefly discussed below:

5.2.1 Basic Configuration

Basic configuration allows to select the target attribute to forecast from dataset attributes. Additionally, the package let user set the necessary parameters such as, Number of units to forecast, Time stamp, Periodicity and Confidence intervals. Figure 5.1 shows the basic configuration panel.

5.2.2 Advanced Configuration

More critical and essential settings are configured in advanced configuration to acute the forecast performance.

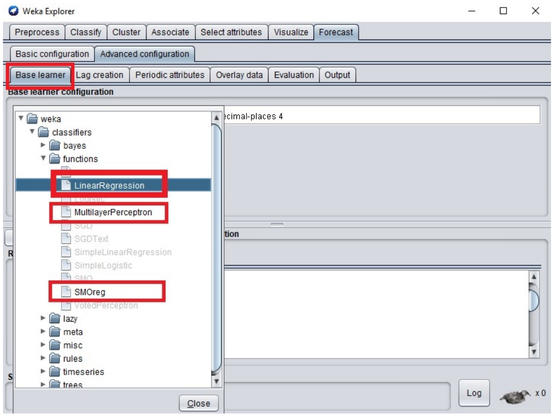


Figure 5.2: Base learner panel in advanced configuration.

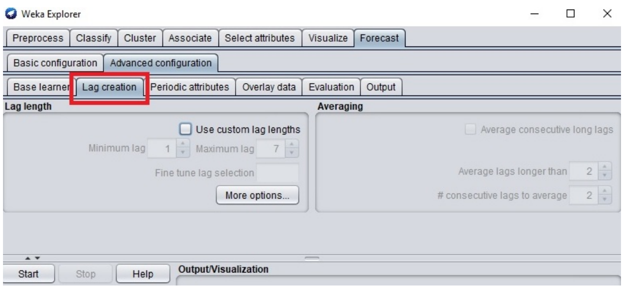


Figure 5.3: Lag creation.

Base Learner

Base learner is the platform to select a particular algorithm to build model based on the time series. Figure 5.2 gives a detail picture of the panel.

Lag Creation

Lag variables define the relationship between the previous and present values. Based on the Periodicity parameter from Basic Configuration some default lagged variable values are generated. The package allows to use custom lag lengths with minimum and maximum lag values. Detail is in figure 5.3.

Periodic Attributes

This platform allows to adjust the date derived variables. The function is available only for the dataset containing datetime stamp.

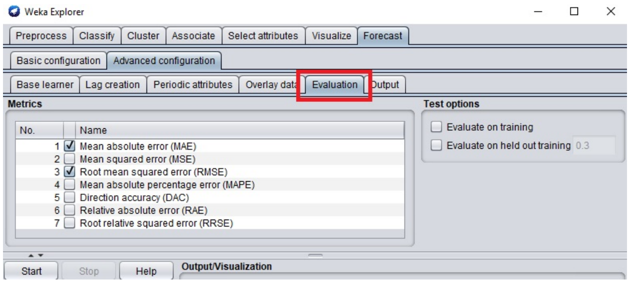


Figure 5.4: Evaluation panel.

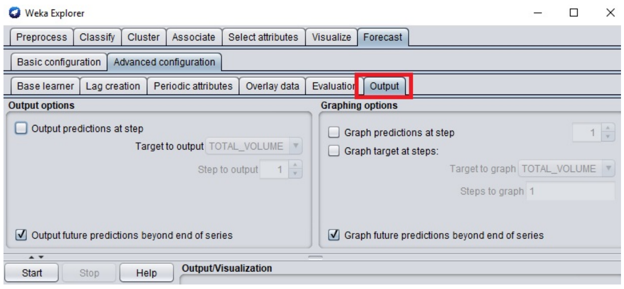


Figure 5.5: Output Panel.

Overlay Data

The data that are not supposed to be forecast can be specified as overlay data. The package by default do not use any overlay data. But, it can be customized for special cases (e.g. random incidents, market crash) as per user's requirement.

Evaluation

The evaluation platform provides a set of metrics (e.g. Mean absolute error, Mean squared error, Root mean squared error, Mean absolute percentage error, Direction accuracy, Relative absolute error and Root relative squared error) to measure the performance of the forecasting model. Figure 5.4 demonstrates the overall system.

Output

The final output can be produced in both textual and graphical representation. Checking the box 'Output prediction at step' outputs the forecasted and real value for any particular single step. Figure 5.5 provides a detail view of the process.

5.3 Experimental Results

Each algorithm mentioned before generated model based on the training data. The study applied these regression models to forecast the pattern of total vehicle flow for the next seven days. Figure 5.1 shows the line graphs predicted by each of the regression models along with the actual traffic flow and the previous historical traffic flow. The actual traffic flow was used to measure the accuracy of the generated models. Figure 5.2 gives a much clear view of the predicted lines along with the actual traffic flow. The study have also compared the performance of the models using mean absolute percentage error (MAPE).

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Actual_i - Predicted_i|}{Actual_i} * 100 \tag{5.1}$$

Here, N = 7, the number of days for prediction, and $Actual_i$ and $Predicted_i$ representing the actual traffic volume and the predicted traffic volume respectively.

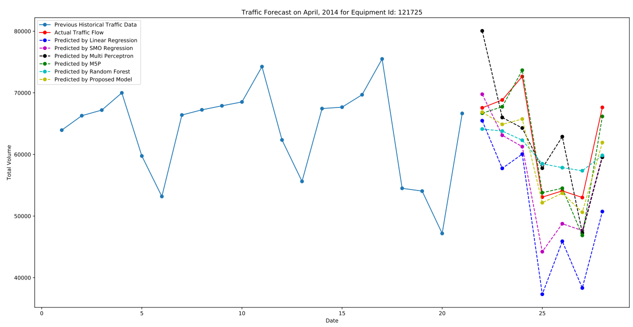


Figure 5.6: Traffic flow forecasting - with historical traffic data.

From Table 5.1 we can see the error rate generated for each of the regression models for a particular road of the city with the equipment id 121725. MAPE is given in the last row of the table for each regression model. This table concludes that M5P based regression tree has the lowest error rate among these regression models. It is clear from the table that even though our proposed model did not show better accuracy than M5P based model tree however, the model has ascertained its efficacy by having the second most lowest error rate. Although, we have shown the prediction accuracy for one road

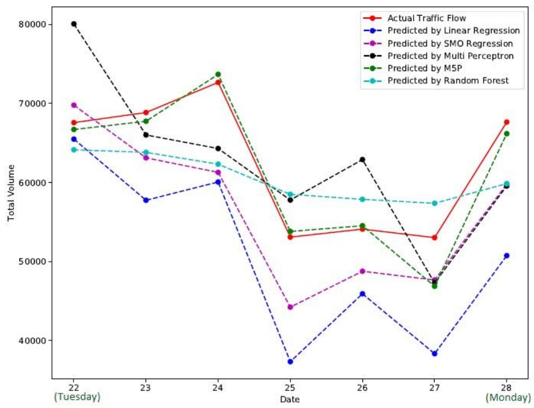


Figure 5.7: Traffic flow forecasting (zoomed in) - without historical traffic data.

of the city, the same approach been applied over other roads of the city as well and M5P resulted better accuracy than other algorithms.

Table 5.1: Result.

Day	Linear Regression	SMO Regression	Multilayer Perceptron	M5P Model Tree	Random Forest	Proposed Model
22nd Apr	3.08%	3.26%	18.48%	1.28%	5.08%	1.03%
23nd Apr	16.12%	8.32%	4.12%	1.60%	7.32%	5.74%
24nd Apr	17.34%	15.67%	11.51%	1.40%	14.24%	9.50%
25nd Apr	29.69%	16.68%	8.84%	1.35%	10.19%	1.71%
26nd Apr	15.12%	9.86%	16.26%	0.78%	6.96%	0.70%
27nd Apr	27.69%	10.10%	10.81%	11.60%	8.19%	4.50%
28nd Apr	24.98%	11.63%	11.95%	2.16%	11.51%	8.43%
MAE	19.15%	10.79%	11.71%	2.88%	9.07%	4.52%

5.4 Summary

Detail summary of the experimental setup and results are discussed in this chapter. The figures shown in this chapter provides a clear view of the experimental result. The results suggest that M5P based regression tree has better accuracy than all the other algorithms.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

It is not far away that intelligent transportation will transform our regular transportation safer, greener and more convenient. ITS can reduce the number of traffic accidents, prevent injuries and severities, ensure better traffic flow and better fuel efficiency. Therefore, we need to come up with profound ideas to make the future transportation infrastructure stronger and more efficient. This study provide a brief survey of traffic flow prediction in recent time and uncover a simple yet effective approach to predict the future traffic flow for Porto city, Portugal. In this paper, we have prepared the traffic data of the month April, 2014 that was collected from different roads using 22 loop sensors. We pre-processed road traffic data to apply regression models for forecasting road traffic flow. This work, took first three weeks traffic data of the month to train the models and test prediction performance of the models for the fourth week. Each model generated patterns to predict the future traffic flow for different roads. We have compared the performance of five different regression models. M5 based regression tree performed the best among them including the ensemble model proposed in the study with the lowest error rate.

6.2 Future Work

Further researches can be done to improve the performance of our approach, such as, by reducing the time interval to produce more meaningful results. Moreover, this approach can even be further used to predict traffic flow for more than just one week, e.g. for a whole month or more. However, it is not obvious that same regression model can give the same accuracy for all over the world. It may vary depending on different circumstances on different roads. But, the approach presented in this study definitely works for predicting the real world road traffic. Although, the proposed ensemble model did not outperform other models but, it should be considered that this kind of ensemble model along with a learning parameter or weight can produce better result in some other environment. In future, we would like to make road traffic visualizing and forecasting model for some of the roads of the capital of Bangladesh, Dhaka city.

Bibliography

[1] I. Alam, M. F. Ahmed, M. Alam, J. Ulisses, D. M. Farid, S. Shatabda, and R. J. F. Rossetti, "Pattern mining from historical traffic big data," in *IEEE Technologies for Smart Cities (TENSYMP 2017)*. IEEE, July 2017, pp. 1–5. 2

[2] F. B. Ghorghi and H. Zhou, "Traffic control devices for deterring wrong-way driving: historical evolution and current practice," *Journal of Traffic and Transportation Engineering*, vol. 4, pp. 280–289, 2017. 2

[3] O. Altintasi, H. T. Yaman, and KaganTuncay, "Detection of urban traffic patterns from floating car data (fcd)," *Transportation Research Procedia*, vol. 22, pp. 382–391, 2017. 2

[4] S. Zhao, P. Zhao, and Y. Cui, "A network centrality measure framework for analyzing urban traffic flow: A case study of wuhan, china," *Physica A: Statistical Mechanics and its Applications*, vol. 478, pp. 143–157, 2017. 2

[5] L. B. Moncayo, J. L. Naranjo, I. P. García, and R. Mosquera, "Neural based contingent valuation of road traffic noise," *Transportation Research Part D: Transport and Environment*, vol. 50, pp. 26–39, 2017. 2

[6] X. Chen, L. Peng, M. Zhang, and W. Li, "A public traffic demand forecast method based on computational experiments," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 4, pp. 984–995, 2017. 2

[7] M. Gaber, A. M. Wahaballa, A. M. Othman, and A. Diab, "Traffic accidents prediction model using fuzzy logic: Aswan desert road case study," *Journal of Engineering Sciences Assiut University*, vol. 45, p. 2844, 2017. 2

[8] W. Hu, L. Yan, H. Wang, B. Du, and D. Tao, "Real-time traffic jams prediction inspired by biham, middleton and levine (bml) model," *Information Sciences*, vol. 381, pp. 209–228, 2017. 2

[9] T. Liebig, N. Piatkowski, C. Bockermann, and K. Morik, "Dynamic route planning with real-time traffic predictions," *Information Systems*, vol. 64, pp. 258–265, 2017. 2

[10] Z. Hou and X. Li, "Repeatability and similarity of freeway traffic flow and long-term prediction under big data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 6, pp. 1786–1796, 2016. 3

[11] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: a deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2015. 6

[12] Y. Duan, Y. Lv, and F.-Y. Wang, "Performance evaluation of the deep learning approach for traffic flow prediction at different times," in *IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI), 2016*. IEEE, 2016, pp. 223–227. 7

[13] T. Zhou, G. Han, X. Xu, Z. Lin, C. Han, Y. Huang, and J. Qin, "δ-agree adaboost stacked autoencoder for short-term traffic flow forecasting," *Neurocomputing*, vol. 247, pp. 31–38, 2017. 7

[14] A. Koesdwiady, R. Soua, and F. Karray, "Improving traffic flow prediction with weather information in connected cars: A deep learning approach," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 9508–9517, 2016. 7

[15] Z. Zhao, W. Chen, X. Wu, P. C. Chen, and J. Liu, "Lstm network: a deep learning approach for short-term traffic forecast," *IET Intelligent Transport Systems*, vol. 11, no. 2, pp. 68–75, 2017. 7

[16] H. Shao and B.-H. Soong, "Traffic flow prediction with long short-term memory networks (lstms)," in *IEEE Region 10 Conference (TENCON), 2016*. IEEE, 2016, pp. 2986–2989. 7

[17] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 1–17, 2017. 7

[18] R. Fu, Z. Zhang, and L. Li, "Using lstm and gru neural network methods for traffic flow prediction," in *Youth Academic Annual Conference of Chinese Association of Automation (YAC)*. IEEE, 2016, pp. 324–328. 7

[19] C. Yanchong, H. Darong, and Z. Ling, "A short-term traffic flow prediction method based on wavelet analysis and neural network," in *Chinese Control and Decision Conference (CCDC), 2016*. IEEE, 2016, pp. 7030–1034. 8

[20] S. do Oh, Y. jin Kim, and J. sun Hong, "Urban traffic flow prediction system using a multifactor pattern recognition model," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2744–2755, 2015. 8

[21] M. Meng, C. Shao, Y. Wong, B. Wang, and H. Li, "A two-stage short-term traffic flow prediction method based on avl and aknn techniques," *Journal of Central South University*, vol. 22, no. 2, pp. 779–86, 2015. 8

[22] M. Bernaś, B. Płaczek, P. Porwik, and T. Pamuła, "Segmentation of vehicle detector data for improved k-nearest neighbours-based traffic flow prediction," *IET intelligent transport systems*, vol. 9, no. 3, pp. 264–274, 2014. 8

[23] D. Xia, H. Li, B. Wang, Y. Li, and Z. Zhang, "A map reduce-based nearest neighbor approach for big-data-driven traffic flow prediction," *IEEE access*, vol. 4, pp. 2920–2934, 2016. 8

[24] P. Dell'Acqua, F. Bellotti, R. Berta, and A. D. Gloria, "Time-aware multivariate nearest neighbor regression methods for traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3393–3402, 2015. 8

[25] H. Hong, W. Huang, X. Xing, X. Zhou, H. Lu, K. Bian, and K. Xie, "Hybrid multi-metric k-nearest neighbor regression for traffic flow prediction," in *IEEE 18th International Conference on Intelligent Transportation Systems (ITSC), 2015*. IEEE, 2015, pp. 2262–2267. 8

[26] J. Zhao and S. Sun, "High-order gaussian process dynamical models for traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 7, pp. 2014–2019, 2016. 8

[27] X. Ling, X. Feng, Z. Chen, Y. Xu, and H. Zheng, "Short-term traffic flow prediction with optimized multi-kernel support vector machine," in *IEEE Congress on Evolutionary Computation (CEC), 2017*. IEEE, 2017, pp. 294–300. 9

[28] Y. Xu, H. Chen, Q.-J. Kong, X. Zhai, and Y. Liu, "Urban traffic flow prediction: a spatio-temporal variable selection-based approach," *Journal of Advanced Transportation*, vol. 50, no. 4, pp. 489–506, 2016. 9

[29] X. Chen, Z. Wei, X. Liu, Y. Cai, Z. Li, and F. Zhao, "Spatiotemporal variable and parameter selection using sparse hybrid genetic algorithm for traffic flow forecasting," *International Journal of Distributed Sensor Networks*, vol. 13, no. 6, p. 1550147717713376, 2017. 9

[30] J. Ahn, E. Ko, and E. Y. Kim, "Highway traffic flow prediction using support vector regression and bayesian classifier," in *International Conference on Big Data and Smart Computing (BigComp), 2016*. IEEE, 2016, pp. 239–244. 9

[31] E. Kim, "Mrf model based real-time traffic flow prediction with support vector regression," *Electronics Letters*, vol. 53, no. 4, pp. 243–245, 2017. 9

[32] S. Sun, C. Zhang, and G. Yu, "A bayesian network approach to traffic flow forecasting," *IEEE Transactions on intelligent transportation systems*, vol. 7, no. 1, pp. 124–132, 2006. 10

[33] X. Zhou, C. Li, Z. Liu, T. H. Luan, Z. Miao, L. Zhu, and L. Xiong, "See the near future: A short-term predictive methodology to traffic load in its," in *IEEE International Conference on Communications (ICC), 2017*. IEEE, 2017, pp. 1–6. 10

[34] Z. Zhu, B. Peng, C. Xiong, and L. Zhang, "Short-term traffic flow prediction with linear conditional gaussian bayesian network," *Journal of Advanced Transportation*, vol. 50, no. 6, pp. 1111–1123, 2016. 10

[35] D. wei Xu, Y. dong Wang, L. min Jia, Y. Qin, and H. hui Dong, "Real-time road traffic state prediction based on arima and kalman filter," *Frontiers of Information Technology & Electronic Engineering*, vol. 18, no. 2, pp. 287–302, 2017. 10

[36] H. Andre and J. L. Ny, "A differentially private ensemble kalman filter for road traffic estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.* IEEE, 2017, pp. 6409–6413. 10

[37] A. Abadi, T. Rajabioun, and P. A. Ioannou, "Traffic flow prediction for road transportation networks with limited traffic data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 653–662, 2015. 10

[38] F. Schimbinschi, L. Moreira-Matias, V. X. Nguyen, and J. Bailey, "Topology-regularized universal vector autoregression for traffic forecasting in large urban areas," *Expert Systems with Applications*, vol. 82, pp. 301–316, 2017. 11

[39] S. Devi and TNeetha, "Machine learning based traffic congestion prediction in a iot based smart city," *International research journal of Engineering and Technology (IRJET)*, vol. 4, pp. 3442–3445, 2017. 11

[40] D. Apronti, K. Ksaibati, K. Gerow, and J. J. Hepner, "Estimating traffic volume on wyoming low volume roads using linear and logistic regression methods," *Journal of traffic and transportation engineering (English edition)*, vol. 3, no. 6, pp. 493–506, 2016. 11

[41] Y. Rajabzadeh, A. H. Rezaie, and H. Amindavar, "Short-term traffic flow prediction using time-varying vasicek model," *Transportation Research Part C: Emerging Technologies*, vol. 74, pp. 168–181, 2017. 11

[42] Z. Ma, G. Luo, and D. Huang, "Short term traffic flow prediction based on on-line sequential extreme learning machine," in *Eighth International Conference on Advanced Computational Intelligence (ICACI), 2016.* IEEE, 2016, pp. 143–149. 11

[43] R. Soua, A. Koesdwiady, and F. Karray, "Big-data-generated traffic flow prediction using deep learning and dempster-shafer theory," in *International Joint Conference on Neural Networks (IJCNN), 2016.* IEEE, 2016, pp. 3195–3202. 11

[44] A. Salamanis, G. Margaritis, D. D. Kehagias, G. Matzoulas, and D. Tzovaras, "Identifying patterns under both normal and abnormal traffic conditions for short-term traffic prediction," *Transportation Research Procedia*, vol. 22, pp. 665–674, 2017. 12

[45] A. Raza and M. Zhong, "Lane-based short-term urban traffic forecasting with ga designed ann and lwr models," *Transportation Research Procedia*, vol. 25, pp. 1430–1443, 2017. 12

[46] A. Csikós, T. Charalambous, H. Farhadi, B. Kulcsár, and H. Wymeersch, "Network traffic flow optimization under performance constraints," *Transportation Research Part C: Emerging Technologies*, vol. 83, pp. 120–133, 2017. 12

[47] S. Amini, I. Gerostathopoulos, and C. Prehofer, "Big data analytics architecture for real-time traffic control," in *5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2017.* IEEE, 2017, pp. 710–715. 12

[48] R. García-Ródenas, M. L. López-García, and M. T. Sánchez-Rico, "An approach to dynamical classification of daily traffic patterns," *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 3, pp. 191–212, 2017. 13

[49] S. Coogan, C. Flores, and P. Varaiya, "Traffic predictive control from low-rank structure," *Transportation Research Part B: Methodological*, vol. 97, pp. 1–22, 2017. 13

[50] M. Zhou, X. Qu, and X. Li, "A recurrent neural network based microscopic car following model to predict traffic oscillation," *Transportation Research Part C: Emerging Technologies*, vol. 84, pp. 245–264, 2017. 13

[51] N. H. N. Mustapha and N. N. N. Wahidah, "Outflow of traffic from the national capital kuala lumpur to the north, south and east coast highways using flow, speed and density relationships," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 3, no. 6, pp. 540–548, 2016. 13

[52] I. Lana, J. D. Ser, and I. I. Olabarrieta, "Understanding daily mobility patterns in urban road networks using traffic flow analytics," in *IEEE/IFIP Network Operations and Management Symposium (NOMS), 2016*. IEEE, 2016, pp. 1157–1162. 13

[53] C. Xia, J. D. Courtney Cochrane and, G. Fan, E. Holmes, M. M. Guirl, P. Murphy, J. Palmer, P. Carter, L. Slivinski *et al.*, "Assimilating eulerian and lagrangian data in traffic-flow models," *Physica D: Nonlinear Phenomena*, vol. 346, pp. 59–72, 2017. 14

[54] A. Talebpour, H. S. Mahmassani, and S. H. Hamdar, "Effect of information availability on stability of traffic flow: Percolation theory approach," *Transportation Research Procedia*, vol. 23, pp. 81–100, 2017. 14

[55] M. Dell'Orco and M. Marinelli, "Modeling the dynamic effect of information on drivers' choice behavior in the context of an advanced traveler information system," *Transportation Research Part C: Emerging Technologies*, vol. 85, pp. 168–183, 2017. 14

[56] C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang, "The promises of big data and small data for travel behavior (aka human mobility) analysis," *Transportation research part C: emerging technologies*, vol. 68, pp. 285–299, 2016. 14

[57] Z. Cheng, M.-Y. Chow, D. Jung, and J. Jeon, "A big data based deep learning approach for vehicle speed prediction," in *IEEE 26th International Symposium on Industrial Electronics (ISIE), 2017*. IEEE, 2017, pp. 389–394. 15

[58] J. Tang, F. Liu, Y. Zou, W. Zhang, and Y. Wang, "An improved fuzzy neural network for traffic speed prediction considering periodic characteristic," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 9, pp. 2340–2350, 2017. 15

[59] C. Song, HeeyunLee, ChangbeomKang, W. Lee, Y. B. Kim, and S. W. Cha, "Traffic speed prediction under weekday using convolutional neural networks concepts," in *IEEE Intelligent Vehicles Symposium (IV), 2017*. IEEE, 2017, pp. 1293–1298. 15

[60] B. Yao, C. Chen, Q. Cao, L. Jin, M. Zhang, H. Zhu, and B. Yu, "Short-term traffic speed prediction for an urban corridor," *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 2, pp. 154–169, 2017. 15

[61] M. Cebecauer, E. Jenelius, and W. Burghout, "Integrated framework for real-time urban network travel time prediction on sparse probe data," *IET Intelligent Transport Systems*, vol. 12, no. 1, pp. 66–74, 2017. 15

[62] G. A. Seber and A. J. Lee, *Linear regression analysis.* John Wiley & Sons, 2012, vol. 329. 26

[63] J. C. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Microsoft Research, Tech. Rep., April 1998. 27

[64] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004. 27

[65] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to the smo algorithm for svm regression," *IEEE transactions on neural networks*, vol. 11, no. 5, pp. 1188–1193, 2000. 27

[66] M. Minsky, S. A. Papert, and L. Bottou, *Perceptrons: An introduction to computational geometry.* MIT press, 2017. 29

[67] T. Rashid, *Make your own neural network.* CreateSpace Independent Publishing Platform, 2016. 29

[68] J. R. Quinlan *et al.*, "Learning with continuous classes," in *5th Australian joint conference on artificial intelligence*, vol. 92. Singapore, 1992, pp. 343–348. 30

[69] Y. Wang and I. H. Witten, "Inducing model trees for continuous classes," in *Proceedings of the Ninth European Conference on Machine Learning*, 1997, pp. 128–137. 30

[70] T. K. Ho, "Random decision forests," in *Document analysis and recognition, 1995., proceedings of the third international conference on*, vol. 1. IEEE, 1995, pp. 278–282. 31

[71] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001. 31

[72] T. Shi and S. Horvath, "Unsupervised learning with random forest predictors," *Journal of Computational and Graphical Statistics*, vol. 15, no. 1, pp. 118–138, 2006. 32

[73] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009. 33

Appendix A

Appendix A: Additional Experiments

A.1 Traffic Forecast on Other Roads of The City

We have experimented the capability of regression models on traffic flow prediction by testing the models on different roads of the city.

A.1.1 Prediction for Road: A (Eq Id: 121726)

Figures and table are given below:

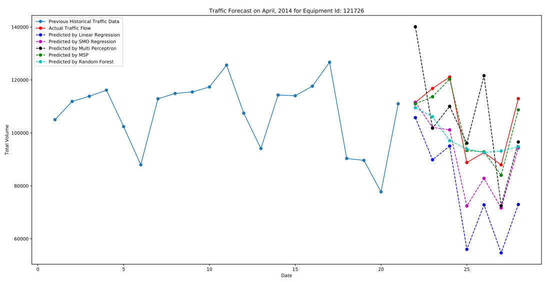


Figure A.1: Traffic flow forecasting for Equipment id: 121726 - with historical traffic data.

A.1.2 Prediction for Road: B (Eq Id: 121735)

Figures and table are given below:

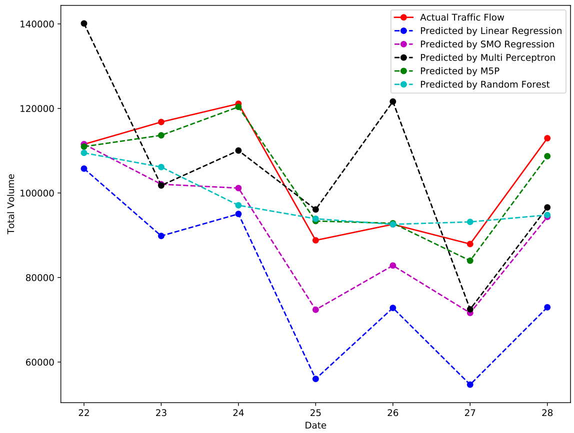


Figure A.2: Traffic flow forecasting for Equipment id: 121726 (zoomed in) - without historical traffic data.

Table A.1: Prediction result for equipment id: 121726

Regression Models	Error Rate (%) Calculation							
	22nd Apr	23rd Apr	24th Apr	25th Apr	26th Apr	27th Apr	28th Apr	MAE
Linear Regression	5.16%	23.07%	21.52%	36.91%	21.36%	37.82%	35.40%	25.89%
SMO Regression	0.025%	12.60%	16.47%	18.48%	10.53%	18.54%	16.48%	13.30%
Multilayer Perceptron	25.62%	12.85%	9.13%	8.21%	31.36%	17.56%	14.48%	17.03%
M5P Model Tree	0.49%	2.69%	0.61%	5.13%	0.26%	4.48%	3.76%	2.49%
Random Forest	1.82%	9.12%	19.83%	5.74%	0.009%	5.94%	16.07%	8.36%

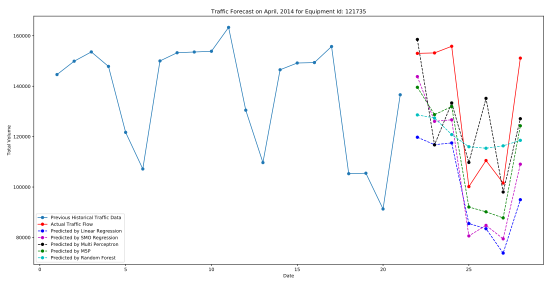


Figure A.3: Traffic flow forecasting for Equipment id: 121735 - with historical traffic data.

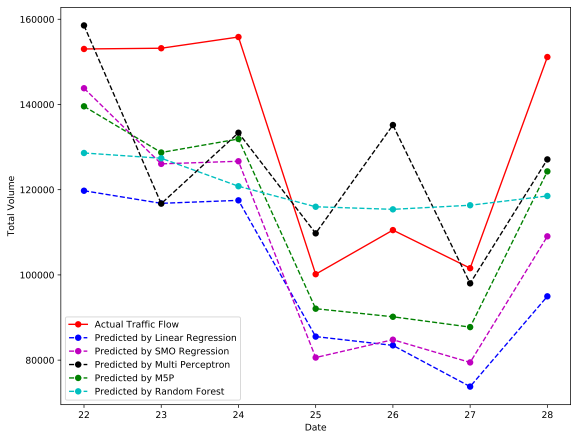


Figure A.4: Traffic flow forecasting for Equipment id: 121735 (zoomed in) - without historical traffic data.

Table A.2: Prediction result for equipment id: 121735

Regression Models	Error Rate (%) Calculation							
	22nd Apr	23rd Apr	24th Apr	25th Apr	26th Apr	27th Apr	28th Apr	MAE
Linear Regression	21.74%	23.75%	24.59%	14.61%	24.47%	27.35%	37.15%	24.81%
SMO Regression	6.02%	17.71%	18.71%	19.52%	23.27%	21.77%	27.84%	19.26%
Multilayer Perceptron	3.60%	23.80%	14.41%	9.57%	22.31%	3.50%	15.89%	13.30%
M5P Model Tree	8.79%	15.97%	15.35%	8.08%	18.40%	13.61%	17.75%	13.99%
Random Forest	15.95%	16.85%	22.47%	15.80%	4.40%	14.53%	21.57%	15.94%

A.1.3 Prediction for Road: C (Eq Id: 121754)

Figures and table are given below:

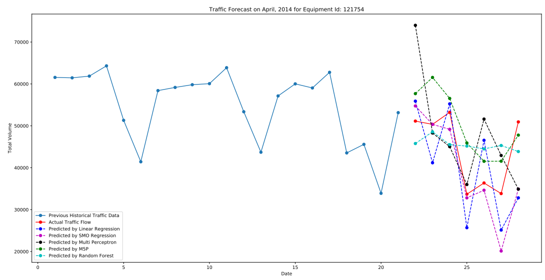


Figure A.5: Traffic flow forecasting for Equipment id: 121754 - with historical traffic data.

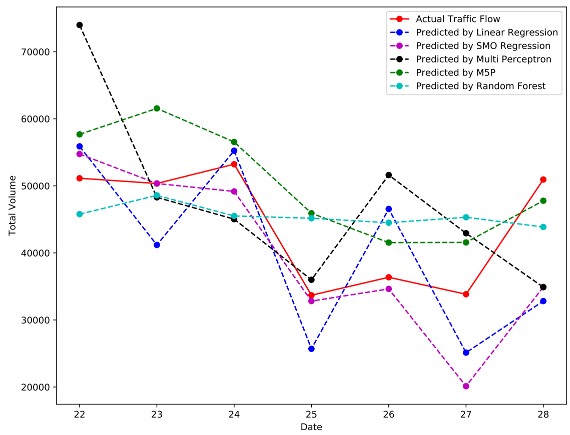


Figure A.6: Traffic flow forecasting for Equipment id: 121754 (zoomed in) - without historical traffic data.

Table A.3: Prediction result for equipment id: 121754

Regression Models	Error Rate (%) Calculation							
	22nd Apr	23rd Apr	24th Apr	25th Apr	26th Apr	27th Apr	28th Apr	MAE
Linear Regression	10.46%	33.64%	14.83%	38.92%	0.41%	37.71%	46.92%	26.13%
SMO Regression	12.29%	18.87%	24.19%	21.99%	25.30%	50.11%	43.53%	28.04%
Multilayer Perceptron	18.52%	22.20%	30.57%	14.44%	11.32%	6.43%	43.56%	21.01%
M5P Model Tree	7.59%	0.83%	12.83%	9.15%	10.40%	3.03%	22.72%	9.51%
Random Forest	26.66%	21.73%	29.84%	7.36%	4.01%	12.32%	29.05%	18.71%

A.1.4 Prediction for Road: D (Eq Id: 121756)

Figures and table are given below:

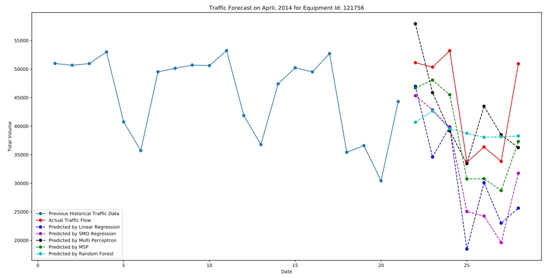


Figure A.7: Traffic flow forecasting for Equipment id: 121756 - with historical traffic data.

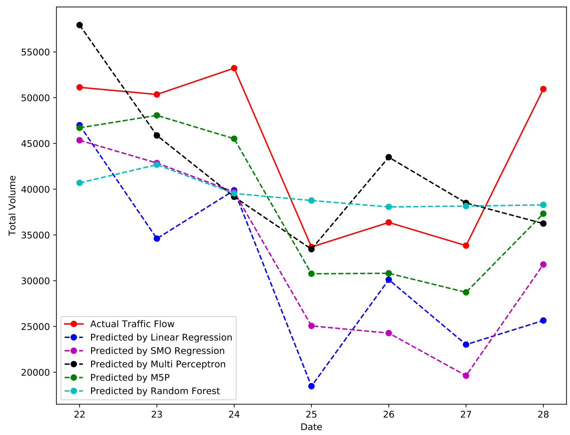


Figure A.8: Traffic flow forecasting for Equipment id: 121756 (zoomed in) - without historical traffic data.

Table A.4: Prediction result for equipment id: 121756

Regression Models	Error Rate (%) Calculation							
	22nd Apr	23rd Apr	24th Apr	25th Apr	26th Apr	27th Apr	28th Apr	MAE
Linear Regression	8.09%	31.28%	25.08%	45.15%	17.20%	31.96%	49.64%	29.77%
SMO Regression	11.31%	14.88%	25.42%	25.60%	33.24%	42.00%	37.62%	27.15%
Multilayer Perceptron	13.30%	8.90%	26.41%	0.71%	19.59%	13.76%	28.85%	15.93%
M5P Model Tree	8.65%	4.54%	14.48%	8.68%	15.28%	15.08%	26.73%	13.35%
Random Forest	20.41%	15.24%	25.72%	15.06%	4.66%	12.75%	24.83%	16.95%