Predictive Urban Traffic Flow Model using Vehicular Big Data

Anshul Mahajan* and Arvinder Kaur

Department of Computer Science and Engineering, Chandigarh University, Mohali -140413, Punjab, India; anshul.mahajan90@gmail.com, arvindercse.cgc@gmail.com

Abstract

Objectives: Intelligent Transportation System (ITS) is raising the interests of the research community in the field of communication. The main objectives of this work are to implement ITS scenario and evaluate the Traffic Density Estimation for different cities for checking the density of vehicles on the different roads and perform a comparative evaluation of traffic densities in the different number of city based scenarios and to increase the accuracy of the density prediction technique for ITS. **Method/Statistical Analysis:** In this manner, an advanced density prediction technique has been proposed for the prediction of density on the road. The proposed technique is based on the number of vehicles on the road, in which vehicle on the map of the three cities at a particular instance of time has been estimated. After estimating the density of vehicle, prediction of that density has been done which shows that there is a maximum density on the lane and in future density on the road has been predicted. **Finding:** A range of values has been generated with the help of a dedicated simulation scenario using SUMO simulator. The computed values have been used for estimating the density of vehicles and the obtained values has been evaluated to compare the densities of vehicles for these three cities. **Improvement:** When we are talking about traffic, then the term comes in our mind is congestion. To avoid that congestion on the roads, we use the Poisson distribution in which vehicle generates after every instance of time. With the help of this distribution, the amount of congestion on the roads can be decreased and this technique is very effective on the hurdle like jam.

Keywords: Density Estimation, ITS, Linear Regression, Poisson Distribution, R Tool, SUMO Simulator, Vehicular Bigdata

1. Introduction

The field of vehicular communication has raised the interests of analysis community within the field of communication and traveller convenience. ITS is a network that connects the vehicles with the aim of safety of driver and supply economical traffic management with the assistance of affiliation protocols. Vehicular system provides numerous services through ITS for increasing the efficiency of road safety through dedicated applications such as collision avoidance between vehicles, lane changing behavior of vehicles using sensors, and automated navigation system that help the vehicles to communicate with other vehicles, and check the vehicle speed on the road. Based on ITS (Figure 1) shows the road network of vehicles on different

lanes in which communication is done by the employment of signals. Some vehicles are moving in identical direction and some in the opposite direction. The figure also shows the continuous movement of vehicles from lane A to lane B, lane C and lane D. At the traffic lights vehicular density is calculated for a variety of vehicles on the lane by installing detectors on the lanes and therefore, the average speed and time taken by the vehicles to cover the distance can be calculated¹. Therefore, the applications are used for minimizing the broadcasting problems that would recognize the traffic density of every lane. The centralized system of traffic light is used for controlling the congestion on the road and generate totally different style of traffic density on the lanes². Some problems occur in estimating density on a road like whenever each vehicle on the road has to

*Author for correspondence



Figure 1. Road Map of different lanes.

determine its neighboring traffic density by using some easy measurements like the amount of neighbors close to it. If the vehicles know that there is road closure or an accident on the road, then the drivers can avoid that route, and move to the other route to save time³. When talking about safety, our main focus is on emergency vehicles like vehicles utilized in department of local government, fire trucks and many other vehicles that work for the organization. In case of emergency, there is more likelihood of accident of those vehicles that are doing jobs for organizations as compared to the others vehicles⁴. The definition of traffic density estimation is based on the idea of traffic flow, which is calculated by the number of vehicles on a specific road⁵. Communication between vehicles is critically important. Vehicles have to send warning messages to alternative vehicles to avoid collision. This information is collected by each vehicle on the road, and each vehicle measures the passing time of vehicles in a given space⁶. An infrastructure-free approach is used for assembling traffic data to avoid the congestion on the roads and a traffic flow model is used to estimate the traffic data. Using the traffic flow model, traffic flow on the road is predicted to avoid congestion, and the new path selected by the vehicle is known as candidate path. Candidate path is the other route for the vehicle to avoid the means wherever there is congestion on the road and vehicle reaches the destination quickly. When communication endpoints do not come under the range of transmission, unicast routing is needed to determine the communication path between vehicles. In this case, some applications that are utilized in the area of inter-vehicle communication need routing algorithms for providing a link between vehicles in the network². When the vehicles are under the range of the transmission then the performance of the traffic density estimation that was based on the vehicle that is near the other vehicle and also used to evaluate the performance of the scheme based on clusters of the vehicles for an organized network⁸. Before estimating the vehicle density on the real scenario, it is important to explore the characteristics of vehicles spacing distribution. Traditionally it was not easy to measure the vehicle space in the traffic on the actual road. Because of this issue, the vehicles spacing distribution are formulated from vehicles point of view². A number of technologies are used for collecting a large amount of vehicular data. Initially, the data was generated by manually, but now this task can be done by machines. In these cases, humans only give commands to machines and machines provide results to human's instructions as a response. Numbers of devices are placed on the vehicle for collecting vehicular data either in small or large amount. Many organizations are still working on better ways to obtain the solution for collecting bigdata effectively. Today, everybody is talking about massive information such as how to store huge information, where to keep the collected information, how much house it needs and how to use that information. Now-a-days the data are large in size that is generated by the government offices, business offices and the social sites like facebook, linkedIn, twitter, etc which produce a bulk of data to be collected examined.

Several authors have used the massive knowledge, conception in varied fields like business, healthcare, education field, industries, government offices and much additional. In every second, the enormous quantity of knowledge is generated from a distinct variety of sources. In vehicular system, multiple techniques are used for assembling knowledge, such as numbers of sensors are placed on the road for managing a giant amount of data and a few sensors on the mobile phones communicate with the vehicles. A set of approaches and techniques have been proposed for grouping vehicles data such as vehicle movement, vehicle id is used to generate emergency warnings. The key idea was based on the several techniques used in infrastructure-free approaches such as vehicleto-vehicle (V2V) communication for estimating density. The authors discussed the different number of strategies, algorithms and methodologies for estimating the amount of the vehicles on the road. A Model based Infrastructurefree Collection and Estimation (MICE), a dynamic path designing for finding the shortest path have been proposed, with minimum time by analyzing and aggregating the vehicle data on the road. The information received from the infrastructure-free and traffic flow model is used

for calculating the speed and density of vehicles on the road that is used for estimating the condition of traffic and hence, analyzing the received data. The performance of the vehicles density estimation was depends on the vehicle that is near to the other vehicle and a scheme that is used for measuring the density is based on clustering that has been used for an organized network¹⁰. Before estimating the vehicle density on the real scenario, it is important to explore the characteristics of vehicles spacing distribution. Traditionally it was not easy to measure the vehicle space in the traffic on the actual road. Because of this issue, the vehicles spacing distribution are formulated from vehicles point of view. A different range of techniques was used for getting information that was derived when processing the condition of traffic within the major cities and alternative extremely packed areas. The sensing technologies were divided into number of devices that were placed on the road or on the vehicle such as road sensors, vehicle sensors, pressure pads, etc. For communication, sensors are linked together with the techniques such as navigation systems that monitor the road conditions such as the location and the average speed of the vehicles. An example of collecting information from different sources such as the records from any social network and also the faculty or school library which is in unorganized format, and that format is not comfortable for humans to grasp it quickly. The information collected from different sources was formatted initially and then employed for analysis. The authors also described about the challenges that occur during collecting information and researching. The information received from the sources was collected and its appropriate structure was created and then, an analysis of the collected information was done in¹¹. Different issues that faced by the individuals and the solution to the challenges that occur due to the weather or any tragedy like floods, landslide, cloudburst, and cyclone was given. If proper pointers and alert messages were send to individuals before calamities, then the damage was reduced. It was facilitated to manage quick helplines like fire trucks, ambulances¹². The information collected within the variety of original trace data of some vehicles counted in one month. The main focus was on the location information within the different kind for every pair of vehicle. All contacts between each pair of vehicle were analyzed and found a distribution like vehicles, inter-contact time, followed by a broad range of the exponential distribution. A quality model was also designed to contain data forwarding protocols, and their analysis performance was done¹³.

Focused on the quantity of problems in the lower layer of the wireless technologies based on vehicular communication spaces. This communication technology includes the transmission power, bandwidth of the communication, its data rate, description of channel etc14. Review the application and the agent technology in ITS system and also discussed the different agents in ITS and their applications using the results and described about improving the performance of traffic additionally as facility¹⁵. A Wireless Sensor Network (WSN) applications in ITS and a new method for the choice of intersection of achieving greater results by deploying sensors and the controllers that are applied at the intersections for increasing the efficiency of traffic in ITS16. Various research has been done on the number of issues when there are a number of paths to the same destination in which some are highly populated and others are low populated and very few people knows about those paths. A system has been developed for tracing the estimated density of the vehicles such as heavy as well as a light vehicle found in different locations at different times, which will help the users to decide the best path to the destination. And also helps the users can also estimate the time that it will take to reach the destination from that path¹⁷. Based on the traffic density measurement on the roads, an intelligent fuzzy controller technique with vision computing is used to manage the traffic lights that have been implemented and an algorithm that is used to solve the problem of road traffic congestion in high congestion hotspots in developing regions¹⁸.

2. Proposed Methodology

Although a number of algorithms for density estimation have been proposed, but data analytics has not been used for estimating vehicles density. In this paper, a density estimation scheme for vehicles has been proposed by considering simulated vehicular traffic in three cities of India. The cities that have been considered are Chandigarh, Delhi and Kolkata. (Figure 2, Figure 3 and Figure 4) depicts the snapshots of roadside scenario for these three cities that have been taken using Simulation of Urban Mobility (SUMO) simulator. The process for density estimation is performed in a number of phases. The main phases of this scheme are Data extraction and integration, Data Analysis and Density estimation. All the phases and the relationship between them are illustrated in (Figure 5). The working of each phase of this process is explained in detail as follows.



Figure 2. Chandigarh City.



Figure 3. Delhi city.



Figure 4. Kolkata city.



Figure 5. Different phases of the process and their attributes.

2.1 Data Extraction and Data Integration

In this phase, data is extracted from the trace file in the form of vehicle and lane information. This generated data are in the form of different structured files such as in formats that makes it complex for manual analysis. In the proposed Density Estimation scheme, a traffic map is generated for urban roadside scenario by considering road maps of the city and vehicles are allowed to enter the lanes from random points after a fixed time interval using Poisson distribution based vehicle generation scheme is used to insert vehicles into the traffic Simulation environment. Each vehicle follows a predetermined traffic map. Different features related to traffic of vehicles such as the position of the Vehicle, Lane ID, Time, Speed are extracted after every simulation. The data integration sub phase combines the extracted data from different sources and provides a unified view of data. In the scenario, data are collected in the form of Open Street Map (OSM) file, once input is given in the form of network file. The network file contains a lot of number of edges and lanes that are linked together. Different identities are assigned to each lane, edge and the file also contains the length of lane. After creation of network file, map file of scenario is checked where data such as vehicle movement, starting and every exiting points of vehicles and a particular timestamp for vehicles, longitude latitude for the vehicles and visibility of the vehicles on the lane are stored in the form of records of every vehicle. A file represents the shape of the objects or vehicles, type of the vehicle, the color of the object is also attached to the map file. Poly file is capable of applying different attributes to the shapes that is imported from sources. Next step is the process of combining route file and additional file (route alternate file), which is used to determine the route of a network. The vehicle routes are computed by the use of SUMO. An additional file used in the route file, provides the definition of the vehicle such as what type of vehicle is used in the scenarios, route definition of vehicle and route distribution for every vehicle. For generation of vehicles on any lane, Poisson Distribution has been employed. In this distribution, the probability with which vehicles are input into the roads over a specific time interval is defined as

$$\mathbf{P}(\mathbf{x}) = \frac{e^{-\mathbf{A}\mathbf{A}\cdot\mathbf{x}}}{\mathbf{x}!} \tag{1}$$

where the symbol (e) represents Euler constant, (λ) represents the mean of the total value and x represents the

total number of vehicles generated over the specified time interval. In the next step, a trip file is created for every vehicle starting from on the lane where the vehicle starts moving and stops according to the trip file. After this a type map file is then created from simulations. This file contains different number of attributes such as Polygon type ID, name, color and category; and these attributes represent the shape of the object, name of the object, the color of the object and which category that object belongs to and the configuration file in which all the process is done by attaching the network file, route file, poly file. The file also contains the beginning and ending time of simulation. At the end of this process, a file is generated in the form of the output file in which all the data is in the combined form and that contains various types of data such as Vehicle ID, Position, Time, Speed, Lane ID. After configuring this files the next task is to separate the data from a file so as to categorize data.

2.2 Data Analysis

In (Figure 3), a sample data where vehicle movement, time to time its position, speed and other attribute that are filtered are represented. With the help of R tool, the attributes are separated from the generated data and represents that filtered data in the tabular form that shows the features of vehicles, lanes, routes, edges. This data is aggregated into CSV format and that represents the sequence of records in which each and every vehicle has its own unique list. After filtering the data, unrequired data need to be discarded and useful information is therefore left in the file. This information is in the form of results derived from the initial step after filtering the data, and that generated data is used for the task. Now the estimation of vehicle movement is performed with respect to sampler vehicle, which samples the vehicle density at specific intervals. The duration of density estimation is uniformly distributed and the value of this duration ranges from 0 to 600 in our scheme. A vehicle that collects all the information is considered as a sampler vehicle. All vehicles that are considered by the sampler for estimation are within the communication range of sampler vehicle. The instantaneous vehicle lane density is initially computed using the following equation:

$$\delta(\mathbf{x}) = \frac{\mathbf{x} = \mathbf{1}}{lg \cdot \mathbf{1}} \tag{2}$$

where Lg denotes the length of the lane, while L denotes the number of lanes and x is the number of vehicles and its position on the lane at any time instance *ti*. Every sampler vehicle helps to compute the number of vehicles within its communication range and also on the same lane as that of the sampler. In the given period of estimation (t, t + T), the value of the average estimated density is obtained using equation (3)

$$\delta^{-} = \frac{\Sigma t_i \in (t, t+1) \delta(tl) (tl+1 \star tl)}{T}$$
(3)

After computing the average density on the lane for each sampler, the next step involves predicting the density of vehicles on the same lane based on previous density. The proposed scheme estimates the density based on Linear Regression. The main reason for using Linear Regression is that it provides a standard performance for predicting the number of vehicles on the lane. This method is also helpful for processing of data easily and faster. This method is also used for predicting the density of the vehicles on a particular lane. Based on the general equation of linear regression technique, the estimation of density can be found out. Equation (6), shows that how big is the contribution of the variable P that is an independent variable towards dependent variable Q. The value of x is determined by the use of the following equation:

$$x = \frac{(\sum Qi) * (Pi)^{2} - (\sum Pi) * (\sum Pi^{*}Qi)}{n^{*}(\sum Pi^{2}) - (\sum Pi)^{2}}$$
(4)

The value of y is determined by the use of following equation:

$$Y = \frac{n \sum Pi * Qi - (\sum) * (\sum Pi Qi)}{n (\sum Pi)^2 - (\sum Pi)^2}$$
(5)

The general equation of simple linear regression is shown below:

$$\mathbf{P} = \mathbf{x} + \mathbf{Q} * \mathbf{y} \tag{6}$$

Where P represents independent variable, Q represents dependent variable, x = P if Q = 0, y represents Gradient, n is the total number of vehicle. From the equation (4) and (5) we easily find out the value of x and y. After putting the values of x and y in equation (6), we get the value of P and Q. For accuracy in the results, we are using the Root Mean Square Error (RMSE) method; this method is helpful for finding the difference between the measured and the estimated prediction. The equation for RMSE is as follows:

RMSE =
$$\left(\frac{1}{n}\sum_{i=0}^{n} |f(pi) - Qi|^2\right)^2$$
 (7)

2.3 Density Estimation

Data generated from the previous stages of the density estimation process cannot be utilized till the last step of this process is executed. By converting the data into a uniform format, the complexity of the density estimation process can be reduced, and thereby increasing the efficiency of the process. The methodology for the scenario of the three cities is represented in (Figure 6). In the scenario, the number of vehicles, edges, routes for the vehicles are considered and connectivity to the network is done by configuring the file. In every simulation, a large number of vehicles are moving in the different number of lanes. Every vehicle is identified by its unique identity number. Every vehicle in the simulated scenario is provided a system generated identification number that is then used for capturing the vehicle description. The features of the vehicles are then used for estimating the maximum and minimum density on the lane. The objectives of the proposed work are estimate the density of the vehicles at a periodic time duration on entering and exiting point of the lane; and to compute average traffic density of the vehicles in the lane of the cities using vehicular big data. After computing average traffic density, then to predict future density using previous density of the cities.



Figure 6. Flow Chart of Proposed Methodology.

3. Procedure

The proposed scheme is implemented using the traffic simulator SUMO¹⁹. In each scenario, the network consists of up to 500 vehicles.

In the simulation, randomly movement of vehicles is calculated after every instance of time. At the rate of 30 vehicles/min enter on the lanes with a high density. In the scenarios, when vehicles reach their destinations, they left the road.

By the use of SUMO simulator, a function called DFROUTER is used for the starting and ending of vehicle movement on the lane. By defining vehicles entering time or exiting from the designed network, three inputs that the DFROUTER takes, such as starting time, exit time and offset. The DFROUTER also helps in selecting the vehicle from the total number of vehicles¹⁹. The simulation starts from 0s and end at 3600s. The parameters that are used in the simulation are summarized in (Table 1).

4. Experimental Results

In this work, the advanced density prediction technique is applied to predict the density on the road. The ITS technique is implemented in SUMO simulator and R tool is used for the analysis of data.

The whole scenario is implemented in SUMO simulator and the analysis data is done by the use of R tool and the graphs between three cities in which it shows a different number of lane changes by the vehicles are presented in (Figure 7). In other words, the figure also shows that the

 Table 1.
 Parameters used in Simulation.

Simulation Parameters	
Number of vehicles	100-500
Simulation time	3600 s
Max Vehicle speed	30 m/s
Length of the lane	5 km
Vehicle entry rate	30 v/min
Number of intersections	15
Number of Lanes	25
Function used for vehicle movement	DFROUTER
Number of Edges	30
Number of routes	17

total number of Lanes changed by the particular vehicle. In city 1 represents the Chandigarh city, city 2 represents Delhi city and city 3 represents Kolkata city. The figure also shows that the movement of vehicles in the different lanes and how much lane passed by the particular vehicle. It also shows the average speed of the vehicle in the lane. The average speed of the vehicle in the lane is calculated by the summation of total number of lanes and the actual speed of the vehicle in the lane divided by the total number of vehicles moved from the lane.

In (Figure 8), number of vehicles passed the total number of lanes at a given instance of time. It also represents how much vehicle is there on the lane at a fixed interval of time and also shows the movement of the vehicle that means in first simulation time how much vehicle move from the particular lane. After every simulation vehicle



Figure 7. Number of lanes change by the vehicles in three cities and average speed of the vehicle in the lane in the cities.



Figure 8. Number of vehicles passed from the total number of lanes at a given instance of time.



Figure 9. Average number of vehicles at main intersection in the three cities and prediction number of vehicles.

can be calculated for Chandigarh city and the same process is done for Delhi and Kolkata city. In (Figure 9), number of vehicles at main intersection in the particular instance of time and also represent the average number of vehicles on the intersection. It also shows that density of vehicle on the lane at particular instance of time and by considering the previous density, estimating the future density on the lane with the same instance of time.

5. Conclusion

In this paper, we have proposed ITS for three cities. The main focus is on the physical properties of vehicles such as vehicle shape, its position, etc. The proposed algorithm for density estimation methods which rely on vehicles communication and information is more suitable for the scalability of ITS. Different starting and end points are selected for vehicles on the different lanes. Numbers of parameters have been used in the simulation for the proposed scheme. Comparison of the three cities is done, which is based on different parameters such as average speed, time taken by the vehicles, number of vehicles on intersections and a number of vehicles on the lane at a particular instance of time. On the basis of these attributes, the comparison of estimated densities between three cities has been done. Based on these estimated densities, future prediction of traffic is done.

6. References

- Rehman US, Khan MA, Zia TA, Zheng L. Vehicular Ad-Hoc Networks (VANETs)-An Overview and Challenges. Journal of Wireless Networking and Communications. 2013; 3(3):29–38.
- 2. Mao R, Mao G. Road traffic density estimation in vehicular networks. In IEEE Wireless Communications and

Networking Conference (WCNC), Shanghai. 2013 Apr. p. 4653–58.

- Zhu H, Fu L, Xue G, Zhu Y, Li M, Ni LM. Recognizing exponential inter-contact time in VANETs. In INFOCOM, Proceedings IEEE San Diego, CA. 2010 Mar; 1–5
- Buchenscheit A, Schaub F, Kargl F, Weber M. A VANETbased emergency vehicle warning system. In IEEE Vehicular Networking Conference (VNC) IEEE, Tokyo. 2009 Oct. p. 1–8.
- Darwish T, Bakar KA. Traffic density estimation in vehicular ad hoc networks: A review. Ad Hoc Networks. 2015 Jan; 24:337–51.
- Kitani T, Shinkawa T, Shibata N, Yasumoto K, Ito M, Higashino T. Efficient vanet-based traffic information sharing using buses on regular routes. InVehicular Technology Conference, VTC Spring IEEE Singapore. 2008 May. p. 3031–36.
- He Z, Cao J, Li T. Mice M. A real-time traffic estimation based vehicular path planning solution using vanets. International Conference on Connected Vehicles and Expo (ICCVE) IEEE, Beijing. 2012 Dec. p. 172–78.
- Song H, Lee HS. A survey on vehicle density estimation in vehicular safety communications and its challenging issues. 16th International IEEE Conference on Intelligent Transportation Systems (ITSC) The Hague. 2013 Oct. p. 2427–32.
- Panichpapiboon S, Pattara-Atikom W. Exploiting wireless communication in vehicle density estimation. IEEE Transactions on Vehicular Technology. 2011 Jul; 60(6):2742–51.
- Rao KV, Ali MA. Survey on big data and applications of real time big data analytics. 2015; 9(8):1–11.

- Tulasi B. Significance of Big Data and analytics in higher education. International Journal of Computer Applications. 2013 Jan; 68(14):1–3.
- Prasad GS, Battula SK, Sruthi CD, Shruthi A, Bhagavathi P, Akhila P. Weather damage prophecy using big-data analytics. International Journal of Advanced Research in Computer Science. 2014 Apr; 5(4):1–1.
- Zhu H, Fu L, Xue G, Zhu Y, Li M, Ni LM. Recognizing exponential inter-contact time in VANETs. In INFOCOM, Proceedings IEEE San Diego, CA. 2010 Mar; 1–5.
- 14. Qu F, Wang FY, Yang L. Intelligent transportation spaces: vehicles, traffic, communications, and beyond. IEEE Communications Magazine. 2010 Nov; 48(11):136–42.
- Chen B, Cheng HH. A review of the applications of agent technology in traffic and transportation systems. IEEE Transactions on Intelligent Transportation Systems. 2010 Jun; 11(2):485–97.
- Khanafer M, Guennoun M, Mouftah HT. WSN architectures for intelligent transportation systems. In 3rd International Conference on New Technologies, Mobility and Security Cairo. 2009 Dec. p. 1–8.
- Goudar R, Singh A, Kumar A. Online Traffic density estimation and vehicle classification management system. Indian Journal of Science and Technology. 2014 Apr; 7(4):508–16.
- Mittal P, Singh Y. Development of intelligent transportation system for improving average moving and waiting time with artificial intelligence. Indian Journal of Science and Technology. 2016; 9(3):1–7.
- Kumar N, Singh JP, Bali RS, Misra S, Ullah S. An intelligent clustering scheme for distributed intrusion detection in vehicular cloud computing. Cluster Computing. 2015 Sep; 18(3):1263–83.