UNIVERSITÉ DE MONTRÉAL

ROAD TRAFFIC CONGESTION ANALYSIS VIA CONNECTED VEHICLES

RANWA AL-MALLAH
DÉPARTEMENT DE GÉNIE INFORMATIQUE ET GÉNIE LOGICIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION
DU DIPLÔME DE PHILOSOPHIÆ DOCTOR
(GÉNIE INFORMATIQUE)
AOÛT 2018

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée :

ROAD TRAFFIC CONGESTION ANALYSIS VIA CONNECTED VEHICLES

présentée par : AL-MALLAH Ranwa
en vue de l'obtention du diplôme de : Philosophiæ Doctor
a été dûment acceptée par le jury d'examen constitué de :

M. PIERRE Samuel, Ph. D., président
M. QUINTERO Alejandro, Doctorat, membre et directeur de recherche
M. FAROOQ Bilal, Ph. D., membre et codirecteur de recherche
Mme BELLAÏCHE Martine, Ph. D., membre
M. ST-HILAIRE Marc, Ph. D., membre

DEDICATION

To my best friend,
you know who you are...

ACKNOWLEDGMENTS

Firstly, Prof. Alejandro Quintero, thank you for believing in me, whitout your support and guidance none of this would have been possible. You were my advisor, my friend. You are a man of few words. This thaught me a lifetime lesson. I learned alot from you. I admire your professionalism and I hope I can be half the teacher you are.

I would like to express my sincere gratitude to Prof. Bilal Farooq for the continuous support of my Ph.D study. You are a genius. You gave me the confidence I need to get to where I am today as a researcher.

I would also like to thank Prof. Martine Bellaïche. You are my idol. I verbally expressed my admiration to you many times and I will do it again here. I will never forget you.

My sincere thanks also go to Prof. Samuel Pierre for being my mentor, not only during my studies, but throughout my career. Every advice you had for me I tried to follow because your words are gold and everyone knows that. Prof. Marc St-Hilaire thank you for showing interest in my work and accepting to participate in this jury.

I also thank my fellow lab colleagues. I honestly wasn't in the lab the person I really am because often times, the thesis required from me to get in my own bubble. I thank you for understanding me.

Last but not the least, I would like to thank mama, baba, Ahmed, Ashraf and Rola. Mama, you raised me and you raised my kids for me. You are madrasti. Baba, I did this for you. I was going to dedicate this to you but you know my best friend and even you, agree that this is dedicated to him. Ahmed, Ashraf, I'm lucky to have you in my life and I am looking forward to enjoy the rest of my life with you. Rola, I love you so much, but this you already know.

RÉSUMÉ

La congestion routière est un état particulier de mobilité où les temps de déplacement augmentent et de plus en plus de temps est passé dans le véhicule. En plus d'être une expérience très stressante pour les conducteurs, la congestion a également un impact négatif sur l'environnement et l'économie. Dans ce contexte, des pressions sont exercées sur les autorités afin qu'elles prennent des mesures décisives pour améliorer le flot du trafic sur le réseau routier. En améliorant le flot, la congestion est réduite et la durée totale de déplacement des véhicules est réduite. D'une part, la congestion routière peut être récurrente, faisant référence à la congestion qui se produit régulièrement. La congestion non récurrente (NRC), quant à elle, dans un réseau urbain, est principalement causée par des incidents, des zones de construction, des événements spéciaux ou des conditions météorologiques défavorables. Les opérateurs d'infrastructure surveillent le trafic sur le réseau mais sont contraints à utiliser le moins de ressources possibles. Cette contrainte implique que l'état du trafic ne peut pas être mesuré partout car il n'est pas réaliste de déployer des équipements sophistiqués pour assurer la collecte précise des données de trafic et la détection en temps réel des événements partout sur le réseau routier. Alors certains emplacements où le flot de trafic doit être amélioré ne sont pas surveillés car ces emplacements varient beaucoup. D'un autre côté, de nombreuses études sur la congestion routière ont été consacrées aux autoroutes plutôt qu'aux régions urbaines, qui sont pourtant beaucoup plus susceptibles d'être surveillées par les autorités de la circulation. De plus, les systèmes actuels de collecte de données de trafic n'incluent pas la possibilité d'enregistrer des informations détaillées sur les événements qui surviennent sur la route, tels que les collisions, les conditions météorologiques défavorables, etc. Aussi, les études proposées dans la littérature ne font que détecter la congestion ; mais ce n'est pas suffisant, nous devrions être en mesure de mieux caractériser l'événement qui en est la cause. Les agences doivent comprendre quelle est la cause qui affecte la variabilité de flot sur leurs installations et dans quelle mesure elles peuvent prendre les actions appropriées pour atténuer la congestion.

Dans cette thèse, nous proposons la collecte de données de trafic via les réseaux ad hoc de véhicules ou VANET. Cette technologie de surveillance avancée est capable d'agréger des variables de trafic microscopiques et macroscopiques à divers niveaux de granularité. Nous avons conçu un algorithme pour la détection et l'évaluation en temps réel de l'état du trafic routier. Nous proposons des modèles de classification basés sur les caractéristiques de trafic collectées pour l'inférence sur la cause de la congestion dans un réseau routier urbain. Nous mettons en place un processus de coopération pour augmenter la précision des estimations

car d'un côté, le trafic est multiforme et aussi pour dissimuler le fait que les véhicules ont une connaissance partielle de l'état de la route. Si les véhicules connectés peuvent détecter la congestion et en attribuer une cause potentielle, nous croyons qu'ils peuvent alors transférer cette connaissance en temps réel à une entité située sur un segment de route en aval pour que cette dernière puisse prédire avec précision le flot sur ce segment. Nous proposons une méthodologie de prédiction de flot de trafic prenant en compte les flots historiques sur le segment en question ainsi que des attributs, tels que les données obtenues en temps réel par les véhicules connectés et les indices de temps de parcours sur les trajectoires des véhicules. Nous montrons comment cette nouvelle approche dans ce domaine améliore la précision de la prédiction. Pour valider les modèles nous simulons des scénarios élaborés à partir de traces réelles de mouvement de véhicules dans un milieu urbain afin de construire un jeu de données synthétique pour le processus d'apprentissage que doivent effectuer les modèles proposés. Nous décrivons dans ce qui suit les trois phases de cette thèse.

La première phase de la thèse affine les approches proposées dans la littérature quant à la détection de la congestion via les réseaux ad hoc de véhicules qui regroupent la congestion non récurrente et la congestion récurrente. Nous soutenons que détecter la congestion ne suffit pas, et nous prévoyons estimer la cause de la congestion, soit étant une congestion récurrente ou non récurrente. Et plus particulièrement dans ce dernier cas, nous estimerons si la cause est due à un incident, zones de construction, événement spécial dans les environs ou à des conditions météorologiques défavorables. Pour ce faire, nous proposons un problème de classification et nous appliquons des méthodes d'apprentissage automatique pour résoudre le problème de classification de la congestion en ses composants en prenant en compte les caractéristiques de trafic collectées à partir des véhicules connectés pour l'inférence sur la cause de la congestion. En particulier, nous considérons un ensemble de caractéristiques uniques pour chaque type de NRC et extrayons ces caractéristiques à partir des données collectées pour déduire la NRC. Plus précisément, les incidents et les zones de construction sont essentiellement caractérisés par des points problématiques sur le segment de route. Pour les conditions météorologiques défavorables, nous évaluons le temps de parcours, la vitesse et la distance inter-véhiculaire tout au long de la trajectoire du véhicule. Et les événements spéciaux sont caractérisés par leur région d'impact et l'accroissement de la demande autour de cette région. Nous intégrons des mécanismes d'apprentissage automatique et des politiques compilées dans les véhicules qui effectuent une détection locale en temps réel pour déduire la cause réelle de la congestion non récurrente. Le classificateur bayésien naïf proposé (NB), l'arbre de décision (CT), la forêt aléatoire (RF) et une technique d'amplification classent avec une grande précision les causes de la congestion détectée. Cette méthodologie peut aider les organismes de transport à réduire la congestion urbaine car sachant les causes sous-jacentes

de la congestion détectée, ils peuvent élaborer des stratégies efficaces pour l'atténuer.

Dans la deuxième phase, afin d'améliorer davantage la précision des estimations de la cause de la congestion obtenue à la phase précédente et d'obtenir des informations en temps réel plus approfondies sur l'état de la circulation, nous présentons des méthodes faisant appel à la coopération décentralisée entre les véhicules connectés. Une méthodologie distribuée basée sur l'exploration de données pour élaborer collectivement une décision concernant la cause de la congestion du trafic sur un réseau routier via les technologies émergentes des véhicules connectés a été développée. Dans l'état actuel, si un événement reçu par un véhicule est une fausse alarme, l'algorithme à bord du véhicule fusionne cette information avec d'autres reçues sur le même segment de route et propage l'incertitude entre les véhicules. Cela pourrait engendrer une congestion encore plus importante. Un processus d'évaluation doit avoir lieu après la détection et avant la fusion des données. Nous ajoutons cette couche pour remédier à la vulnérabilité des algorithmes de fusion et pour réduire les effets secondaires des fausses alarmes car les approches proposées dans la littérature ne traitent pas les données avant de les fusionner. Elles représentent ainsi une menace de sécurité pour le réseau routier. En outre, nous explorons les données collectées à des fins d'apprentissage en construisant des modèles capables d'apprentissage automatique. Nos méthodes d'exploration de données consistent en une procédure de vote, des fonctions de Croyance et une technique d'association de données pour une inférence efficace sur la cause de la congestion du trafic via la technologie des réseaux ah hoc de véhicules. L'évaluation des performances de nos méthodes montre qu'elles améliorent la précision de l'estimation de la cause de la congestion, réduisent le temps de détection et diminuent les fausses alarmes déclenchées dans le réseau. Ceci certifie que les phénomènes complexes de trafic routier sont mieux observés à travers les interactions entre les véhicules échangeant des messages entre eux. Enfin, les simulations démontrent que les méthodes requiert seulement 63% de taux de pénétration de la technologie des véhicules connectés pour obtenir tous les avantages des communications entre les véhicules.

Dans la dernière phase de la thèse, nous abordons le problème de la prédiction du flot de véhicules sur un segment de route donné. Nous intégrons à la prédiction du flot, le fait que les véhicules peuvent détecter une congestion excessive tout au long de leur trajectoire et en attribuer collectivement une cause. Particulièrement, nous incorporons l'impact de divers événements survenant sur la route dans la prédiction du flot de trafic. Nous proposons un réseau de neurones profonds (DNN) et abordons le problème en apprenant le DNN cible dans une technique d'apprentissage multitâche. Les entrées du DNN prennent en compte à la fois les variables macroscopiques et microscopiques du trafic. En effet, en plus des données de flots historiquement observés sur le segment, les données provenant des réseaux ad hoc

de véhicules, tel que l'indice sommaire de temps de parcours éprouvé au long de la trajectoire et les évènements en temps réel vécus sur le réseau urbain sont utilisées par la modèle pour l'apprentissage. Le modèle apprend une représentation prenant en compte les différents événements rencontrés sur les différents segments de sa trajectoire. Les résultats montrent que notre approche surpasse significativement les approches existantes qui ne s'adaptent pas à des situations changeantes de trafic. Le modèle DNN a appris des similitudes historiques entre les différents segments, contrairement à l'utilisation des tendances historiques directes dans la mesure elle-même, car parfois les tendances peuvent ne pas exister dans la mesure, mais le sont dans les similitudes.

En somme, un système de transport est un réseau fortement corrélé. Les caractéristiques des systèmes de transport, tels les grandes quantités de données et les dimensions élevées des variables de la circulation, font de l'apprentissage automatique une approche prometteuse pour la recherche sur les problématiques dans ce domaine. Particulièrement, la prédiction du flot de véhicules permet une modélisation avancée, car la connaissance du volume de trafic allant vers une destination donnera plus d'informations sur les demandes attendues dans un proche avenir. Les techniques proposées pour la collecte de données via la technologie des véhicules connectés, la classification coopérative de la cause de la congestion et la méthodologie développée pour la prédiction du flot aideront les autorités à améliorer le flot de trafic du réseau routier et ainsi réduire la congestion.

ABSTRACT

Road traffic congestion is a particular state of mobility where travel times increase and more and more time is spent in vehicles. Apart from being a quite-stressful experience for drivers, congestion also has a negative impact on the environment and the economy. In this context, there is pressure on the authorities to take decisive actions to improve the network traffic flow. By improving network flow, congestion is reduced and the total travel time of vehicles is decreased. In fact, congestion can be classified as recurrent and non-recurrent (NRC). Recurrent congestion refers to congestion that happens on a regular basis. Non-recurrent congestion in an urban network is mainly caused by incidents, workzones, special events and adverse weather. Infrastructure operators monitor traffic on the network while using the least possible resources. Thus, traffic state cannot be directly measured everywhere on the traffic road network. But the location where traffic flow needs to be improved varies highly and certainly, deploying highly sophisticated equipment to ensure the accurate estimation of traffic flows and timely detection of events everywhere on the road network is not feasible. Also, many studies have been devoted to highways rather than highly congested urban regions which are intricate, complex networks and far more likely to be monitored by the traffic authorities. Moreover, current traffic data collection systems do not incorporate the ability of registring detailed information on the altering events happening on the road, such as vehicle crashes, adverse weather, etc. Operators require external data sources to retireve this information in real time. Current methods only detect congestion but it's not enough, we should be able to better characterize the event causing it. Agencies need to understand what is the cause affecting variability on their facilities and to what degree so that they can take the appropriate action to mitigate congestion.

In this thesis, to optimize the traffic flow in the transportation system in order to mitigate congestion, we propose the collection of measurable traffic features extracted by an advanced monitoring technology, Vehicular Ad hoc NETworks (VANET), capable of aggregating microscopic and macroscopic traffic variables at various levels of granularity. We designed an algorithm for the real-time assessment and evaluation of road traffic condition. We propose classification models based on the traffic features collected for inference on the cause of congestion in an urban road network. We implement a cooperation process to increase estimation accuracy because traffic is multifaceted and to conceal the fact that individually, vehicles have partial knowledge about the road condition. If connected vehicles can detect congestion and cooperatively attribute a possible cause to it, we believe that they can then transfer this knowledge in real time to an entity able to accurately predict flow on a road

segment. We propose a traffic flow prediction framework taking into account historical flows as well as innovative features, such as real-time reports from connected vehicles and travel time along a trajectory for accurate forecasting of flow in an urban network in order to cope with the fact that existing approaches that do not adapt to the varying traffic situations. We show how this novel approach in this domain improves accuracy of prediction. To validate the models and the framework, we simulate scenarios extended from a realistic urban city vehicular motion traces in order to build a synthetic dataset to feed the models for learning purposes. The work in this thesis is carried out in three phases.

In the first phase of our thesis, we refine previous VANET-based congestion detection approaches that group non-recurrent congestion together with recurrent congestion. Not only we propose that detecting congestion is not enough, we plan to further classify the recurrent and non-recurrent congestion (incidents, workzones, special events and adverse weather). We portray this as a classification problem and we apply machine learning methods to solve the classification of congestion into its components taking traffic features collected from connected vehicles into account for the inference on the cause of congestion. Particularly, we consider a set of unique features for each type of NRC and extract such features from the data to infer the NRC. Specifically, incidents and workzones are essentially characterized by problematic spots. For inclement weather, we assess the trajectory travel time, speed and gap. And special events are characterised by their impact region and demand surge. We embed reasoning machinery or compiled policies in vehicles that perform local, real-time sensing to infer the actual cause of the non-recurrent congestion. The proposed Naive Bayesian classifier (NB), Classification Tree (CT), Random Forest (RF) and a boosting technique classify with high accuracy the causes of the underlying congestion status. This framework can assist transportation agencies in reducing urban congestion by developing effective congestion mitigation strategies knowing the root causes of congestion.

In the second phase, to obtain deeper real-time insights of traffic conditions and improve estimation accuracy, we present methods using decentralized cooperation between individual vehicles. A distributed data mining based methodology to elaborate a decision collectively concerning the cause of traffic congestion on a road network via emerging connected vehicle technologies was developed. In the current state, if an event received by a vehicle is a false alarm, the algorithm will fuse the obtained information with others located on a same road segment and spread uncertainty among vehicles and this in turn causes more congestion. An evaluation process has to take place after data sensing and before data fusion. We add this layer to address the vulnerability of fusion algorithms and to lower the side effects of false alarms because the approaches proposed in the literature fail to process the data before fusion

and present a security threat to the network. Also, we explore the collected data for learning purposes by building models capable of machine learning. Our mining methods consist of a voting procedure, belief functions and a data association technique for efficient inference on the cause of traffic congestion via connected vehicles technology. The performance evaluation of the our methods show that they enhance estimation accuracy, lower detection time and decrease false alarms triggered by the network. This implies that the complex traffic phenomena is better observed through the interactions between vehicles exchanging messages. Finally, the methods require only 63% penetration rate to obtain the full benefits of vehicle-to-vehicle communications.

In the last phase of our work in this thesis, we address the problem of traffic flow prediction. We integrate the fact that vehicles traveling along a trajectory can detect excessive congestion and collectively attribute a cause to it into the forecasting of traffic flow on a target road segment. This means that we incorporate the impact of various events happening on the road into the forecasting of traffic flow on a target road segment. We propose a Deep Neural Networks (DNNs), and tackle the problem by learning the target DNN in a multitask learning technique. The DNN input features take into account both macroscopic and microscopic traffic variables in the prediction of traffic flow. In fact, using historical flows and well engineered features, such as real-time reports from connected vehicles and travel time along a trajectory for accurate forecasting of flow in an urban network, the model learns a representation that takes into account the various events that vehicles realistically encounter on the segments along their trajectory. The results show our approach significantly outperforms existing approaches that do not adapt to the varying traffic situations. DNN learned historical similarities between road segments, in contrast to using direct historical trends in the measure itself, since sometimes trends may not exist in the measure but do in the similarities.

In general, a transportation system is a highly correlated network. The characteristics of transportation systems, such as the large amounts of data and the high dimensions of features, makes machine learning a promising approach for transportation research. In fact, traffic flow prediction allows advanced modelling because knowing the volume of traffic heading toward a destination will give more insights about the expected demands in the near future. The proposed techniques for data collection via connected vehicles technology, the cooperative classification of the cause of congestion and the developed framework for flow prediction will help infrastructure authorities improve the network traffic flow and thus reduce traffic congestion.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ANN	Artificial Neural Network
ARIMA	AutoRegressive Integrated Moving Average
BF	Belief Functions
BP	Back-Propagation
CT	Classification Tree
CVs	Connected Vehicles
DAT	Data Assocition Technique
DNN	Deep Neural Network
GPS	Global Positionibg System
ITS	Intelligent Transportation Systems
LJT	Link Journey Times
MLP	Multi Layer Perceptron
MTL	MultiTask Learning
NB	Naive Bayesian
NRC	Non Recurrent Congestion
ns	Network Simulator
Pspot	Problematic spot
RF	Random Forest
RMSE	Root-Mean-Squared Error
RP	extraordinary event ResPonse
RQ	extraordinary event ReQuest
RSU	Road Side Units
SUMO	Simulation of Urban MObility
TAPAS	Travel and Activity PAtterns Simulation
TIS	Traffic Information Systems
V2I	Vehicle-to-Infrastructure
V2V	Vehicle-to-Vehicle
VANET	Vehicular Ad hoc NETworks
VP	Voting Procedure

CHAPTER 1 INTRODUCTION

With the increasing number of vehicles and limited expansion of paved roads, traffic congestion is to be expected. Road traffic congestion happens gradually as vehicles accumulate on a common path. Congestion is a particular state of mobility where travel times increase and more and more time is spent in vehicles. Apart from being a quite-stressful experience for drivers, congestion also have a negative impact on the environment and the economy. In this context, as the complexity of traffic increases, there is pressure on the authorities to take decisive actions to improve the network traffic flow. These actions include optimizing traffic elements such as traffic lights and turning restrictions, lane control, signal timing and route planning. By improving network flow, congestion is reduced and the total travel time of vehicles is decreased.

However, the location where traffic flow needs to be improved varies highly. Since the traffic state cannot be directly measured everywhere, infrastructure operators interpolate information from incomplete, noisy and local traffic data. They are strained to monitor traffic on the road network while using the least possible resources. To scale to larger cities, advanced monitoring techniques should be deployed and must be capable of aggregating traffic data feeds from various levels and at various levels of granularity. Also, the duration and timing of traffic events varies a lot making it difficult to monitor traffic in real time with the conventional monitoring mechanisms. To evaluate the traffic state in real time, operators necessitate extensive data sources to guarantee the accurate evaluation of the traffic state. Besides, well-tailored data sources may not always be available for a particular area of the traffic network. We see that future systems should enable continuous monitoring of the traffic condition along all roads of the traffic network.

One of the hot topics in Intelligent Transportation Systems (ITS) is the development of distributed Traffic Information Systems (TIS) [1] . Such distributed systems monitor and collect data from many sources. These data provide enough comprehensive information in order to better characterize the events detected. Current techniques fail to process the knowledge acquired from the data. In the big data era, techniques should be implemented to make use of the acquired information. Agencies need to understand what is the cause affecting variability on their facilities and to what degree so that they can take the appropriate action to mitigate congestion. Furthermore, at the current stage, the ITS is partially efficient since the vehicle is the only entity that is not contributing to the system. In fact, presently, vehicles are uninformative as they are not engaged in the process of traffic event detection. However,

equipped with a communication technology, vehicles can exchange information and cooperate collectively so as to provide their input to the system. Models should be implemented to make use of the cooperation between vehicles to better assess the road traffic condition. Finally, knowing the volume of traffic heading toward a destination will give more insights about the expected demands in the near future. Indeed, traffic flow prediction allows advanced modelling so that traffic managers take early actions to control the traffic flow and prevent the congestion state. However, current models need to be improve so as to allow fast and more accurate prediction.

In this thesis, to optimize the traffic flow in the transportation system in order to mitigate congestion, we propose the real-time distributed detection and classification of the components of congestion in urban traffic using connected vehicles. Via this next generation sensing technology, we are interested in identifying road traffic events on the basis of exchanging traffic flow data between vehicles. If connected vehicles can detect congestion and cooperatively attribute a possible cause to it, we believe that they can then transfer this knowledge in real time to an entity able to accurately predict flow on a road segment. The traffic flow prediction framework we introduce aims at evaluating anticipated traffic flow at future time frames on a target road segment based on real time feeds provided by connected vehicles and historical data. We show how this novel approach in this domain improves accuracy of prediction so that in real time, valuable information with regard to the potential impacts of the predicted flow can be disseminated to individual drivers and traffic management centers can apply proactive strategies for recovering traffic conditions back to normality. The methods and models we proposed to solve the problems in this thesis are based on machine learning approaches. Since the transportation system is a highly correlated network, with characteristics such as large amounts of data and high dimensions of features, we show how artificial intelligence makes a promising approach for transportation research.

This chapter is divided as follows. Firstly, definitions and basic concepts related to our research are defined and explained to allow for a better understanding of the foundations of our research problem. Then, the addressed problem is described and research objectives are defined. Afterwards, we present the main research contributions and their originality. Finally, the structure of the thesis is outlined.

1.1 Definitions and basic concepts

This section aims at defining the terminology and concepts that will be used in the rest of the thesis, which will help the reader in better grasping the context of our work. We start by introducing the traffic flow theory and describing the microscopic and macroscopic

traffic variables. Then, the data collection methods are explained. Finally, a description of congestion and its components is elaborated.

1.1.1 Traffic flow Theory

In this section, we review the basic concepts related to the traffic flow theory. In transportation engineering, traffic flow is the study of the interactions between travellers, their vehicle and the infrastructure with the aim of understanding and developing an optimal transport network with efficient movement of traffic and minimal traffic congestion problems. Fig. 1.1 shows the interactions between the three elements of the road traffic network.

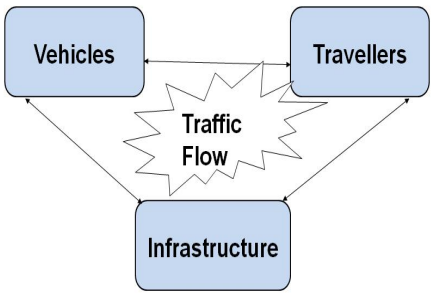

Figure 1.1 Road traffic network

The infrastructure in this context is the road with signage and traffic control devices. It has a quasi-stable geometric configuration. The density is calculated in terms of the number of vehicles per kilometer of road. We also count the number of lanes on a road and classify the infrastructure in terms of highway, road, street segments, lane, junction. The vehicles are of different sizes and are equipped with more and more powerful devices integrating several technologies. Vehicle embedded systems include systems for collecting, processing and disseminating information in the perimeter of the vehicle. The multiple sensors, speedometer, wheel rotation sensors, rain sensors, reversing radars, Global Positioning System (GPS) and mobile phone which tends to be more and more connected to the vehicle by means of Bluetooth technology, the systems of detection by internal and external cameras, positioning sensors on the roadway, obstacle detection radars, they all provide information about the vehicle and its surroundings. Finally, the travellers is the key component of the road system. The traveller usually has a traveling purpose, which is to get from an origin to a destination in a certain period of time. The road traffic network is composed of segments. Each vehicle travels on the road traffic network along a trajectory composed of segments to get from an origin to a destination. It is only when these three elements of infrastructure, vehicles and

travellers are put together that there is traffic flow. Thus, traffic flow prediction approaches cannot be developed without an understanding of traffic flow theory.

Traffic flow theory is a recent field of transportation. It's the characterization of flow through the laws of physics and mathematics. A general, unified and coherent theory has not yet been developed, and the problem has been tackled in different ways. There are different representations of the traffic flow. In what follows we expose the microscopic representation, the macroscopic representation and the variables associated with each representation.

1.1.1.1 Microscopic representation

At the most basic scale of observation, every vehicle is considered individually in the microscopic representation. The following Fig. 1.2 shows the trajectories of eight vehicles on a space-time diagram [2].

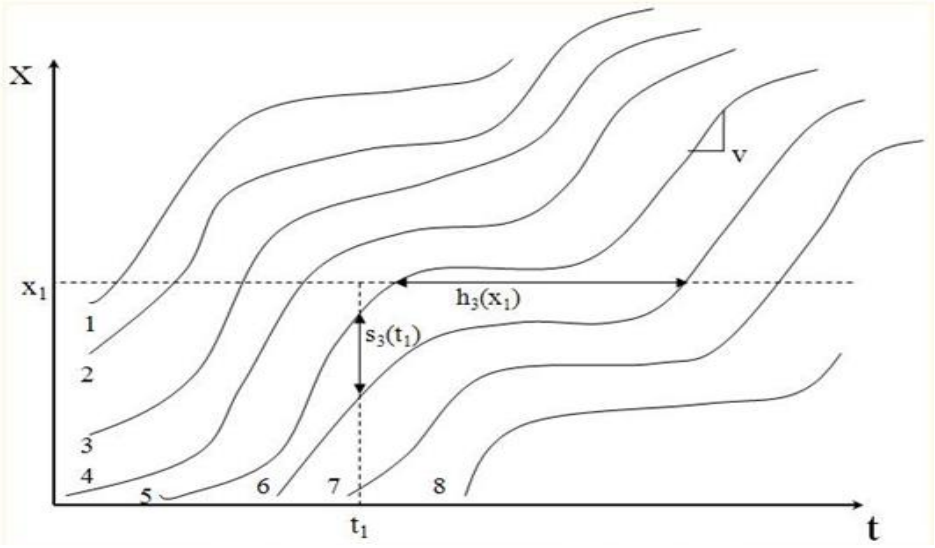

Figure 1.2 Trajectories of eight vehicles on a space-time diagram

We describle below some microscopic variables of the circulation illustrated in the figure :

— v : instantaneous speed, speed has the dimensions of distance divided by time, it is the rate of change of the vehicle's position.
— s : spacing, space separating the front of two successive vehicles at a given instant on the same lane, spacing and instantaneous speed are connected, *s3 (t1)* on the figure.
— h : headway, duration of time between the passage of the front of two successive vehicles at a given point on the same lane. Headway is used to calculate the density of

vehicles. Headway is considered dangerous if *h (x) <1 second*. Traffic can be described in terms of headway, *h3(x1) = s3(t1) / v*.

— Travel time : time required to travel one unit of distance.

— Gap : distance between the rear of a vehicle and the front of the vehicle following it.

1.1.1.2 Macroscopic representation

The macroscopic scale represents the flows of vehicles at a high level of aggregation. It neglects everything that is not average behaviour. Macroscopic models can help characterize different states of the circulation from free flow to congested and cover larger areas, from a road segment to the whole road traffic network. There is a fundamental equilibrium relationship that connects the macroscopic variables. The three variables of the fundamental relationship are density, velocity and flow. We describle them here :

— u : average speed (km/h), average of all distance travelled by each vehicle divided by the duration of the time.

— q : flow (V/h, number of vehicles over time and can be represented by the average flow rate *q(t1, t2, x)* at the abscissa x between the instants *t1* and *t2* which is the ratio *n(t1, t2, x)* of the number of vehicles that passed by x between the two instants.

— k : Density (V/km), distribution of vehicles in space. The link with the microscopic representation is that the average density is the inverse of the mean inter-vehicular distance or spacing.

The fundamental equilibrium relationship is : *q (x, t) = u * k (x, t)* . It's valid when all vehicles move at the same mean spatial velocity and its variables vary simultaneously as shown in Fig. 1.3 [2].

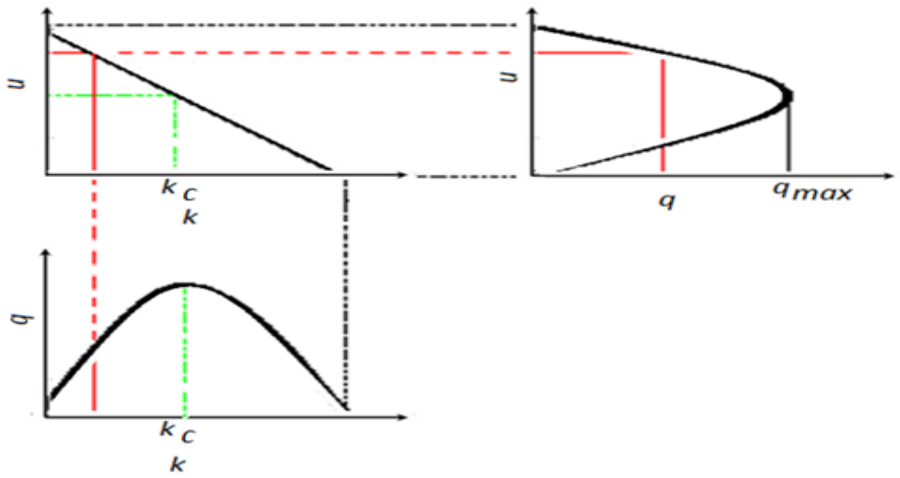

Figure 1.3 Flow-Speed-Density Relationship Diagrams

1.1.2 Traffic data collection

In this section, we present the concept of traffic data collection which refer to the monitoring of traffic by heterogeneous road equipments that measure and collect traffic variables such as traffic gap, density, speed, flow, etc. Subsequently, these data feeds are fused and aggregated to extract useful traffic information. This acquired knowledge from the processed data is used to compute optimal routes for the vehicles, short-term traffic forecasts to reduce road traffic congestion, improve response time to incidents, and ensure a better travel experience for commuters. Finally, a traffic management system can delivers this knowledge to the end users. Traffic data collection is an important phase of the traffic management system and should be done at short intervals to provide good quality because stale data is useless in dynamic environments. In transportation engineering, they rely mainly on the traditional methods based on the infrastructure, such as induction loops, cameras or sensors, to collect the macroscopic or microscopic traffic variables. On the other hand, collection methods done by the vehicles, are floating car data and recently using connected vehicles technologies. We present the traditional methods of collection and connected vehicles technology approach.

1.1.2.1 Traditional methods

Traffic data can be collected from fixed monitoring equipment, such as induction loops, sensors and cameras. Examples of infrastructure sensors that are in-roadway include inductive-loop detectors, which are sawcut into the pavement ; magnetometers, which may be placed underneath a paved roadway or bridge structure ; and tape switches, which are mounted on

the roadway surface. Examples of over-roadway sensors are video image processors that utilize cameras mounted on tall poles adjacent to the roadway or traffic signal mast arms over the roadway ; microwave radar, ultrasonic, and passive infrared sensors mounted in a similar manner ; and laser radar sensors mounted on structures that span the lanes to be monitored [3].

While single inductive-loop detectors give direct information concerning vehicle passage and presence, gap, heading and spacing, other traffic flow parameters such as density and speed must be inferred from algorithms that interpret or analyze the measured data. When these parameters are calculated from inductive-loop data, the values may not have sufficient accuracy for some applications (such as rapid freeway incident detection) or the available information may be inadequate to support the application (such as calculation of link travel time). Furthermore, the operation of inductive-loop detectors is degraded by pavement deterioration, improper installation, and weather-related effects. Street and utility repair may also impair loop integrity. Thus, a good loop installation, acceptance testing, repair, and maintenance program is required to maintain the operational status of an inductive-loop-based vehicle detection system.

Evaluation of over-roadway sensors show that they provide an alternative to inductive-loop detectors. Particularly, video image processing automatically analyze the scene of interest and extract information for traffic surveillance and management. Cameras can replace several in-ground inductive loops, provide detection of vehicles across several lanes. They can classify vehicles by their length and report vehicle presence, density, lane occupancy, and speed for each class and lane. They can track vehicles and may also have the capability to register turning movements and lane changes. Vehicle density, link travel time, and origin-destination pairs are potential traffic parameters that can be obtained by analyzing data from a series of image processors installed along a section of roadway. However, installation and maintenance, include periodic lens cleaning, require lane closure when camera is mounted over roadway. Their performance is affected by inclement weather such as fog, rain, and snow ; vehicle shadows ; vehicle projection into adjacent lanes ; occlusion ; day-to-night transition ; vehicle/road contrast ; and water, salt grime, icicles, and cobwebs on camera lens also, reliable night time signal actuation requires street lighting

In sum, although the traffic flow parameters measured with over-roadway sensors satisfy the accuracy requirements of many applications, infrastructure-based detectors provide fixed-points and short-section traffic information that is extracted from vehicles passing through the detection zone only. The traffic evaluation is restricted to surrounding locations that are close to these installed sensors. Moreover, it is expensive to install and to regularly maintain

these sensor equipments, especially in large downtown scenarios

On the other hand, monitoring can be done using mobile data sources such as GPS-based systems, floating car data of probe vehicle, which are methods based on the vehicle. Probe vehicles or mobile sensors appeared as a complementary solution to fixed sensors for increasing coverage areas and accuracy without requiring expensive infrastructure investment. Two popular types of mobile sensors are GPS-based and cellular-based. GPS-based sensors are sensors with GPS capability and cellular-based sensors are sensors that use information from cellular networks as traffic sensors. Cellular-based sensors are low in cost due to the large number of mobile phones and their associated infrastructures already in service. This potential source of traffic flow data is from cellular telephone companies who monitor the transmitting status of telephones that are engaged in conversations in support of the wireless enhanced all automatic location. The location of these telephones can potentially be made available to traffic management agencies and can assist in estimating congestion and travel time over wide areas. However, most of the probe vehicle techniques that are used for determining the link travel time make use of GPS technology. GPS-based sensors are far more efficient to pinpoint vehicle locations ; thus they can provide highly accurate vehicle movement information. But the major problem with GPS is that the accuracy of a typical GPS receiver is about 10 meters. This makes it difficult to pin-point a crossing for the purpose of congestion measurement. Secondly, it has been noticed that GPS sends erroneous velocity data even when the vehicle is stationary.

New technologies can be used to improve the accuracy, timeliness, and cost efficiency of data collection. In fact, researchers have been focusing their efforts on exploiting the advances in sensing, communication, and dynamic adaptive technologies to efficiently monitor the evolving critical road infrastructure [4], we present the Connected Vehicles (CVs) technology in the next section.

1.1.2.2 Connected Vehicles

Recently, the Intelligent Transportation System (ITS) research has shifted its focus to the next generation sensing technology, Vehicular Ad-hoc NETwork (VANET). The application of wireless technology to moving vehicles enables the creation of vehicular ad hoc networks, also called Connected Vehicles (CVs). Advances in Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) wireless communications have increased the potential of real-time monitoring of traffic variables, for instance, in a distributed manner. Real-time distributed monitoring refers to the process by which macroscopic and microscopic traffic variables are collected by vehicles themselves without the need to send information to a traffic management

center. V2I refers to the communication between the vehicle and the Road Side Unit (RSU). Fig. 1.4 shows the interactions in a connected environment.

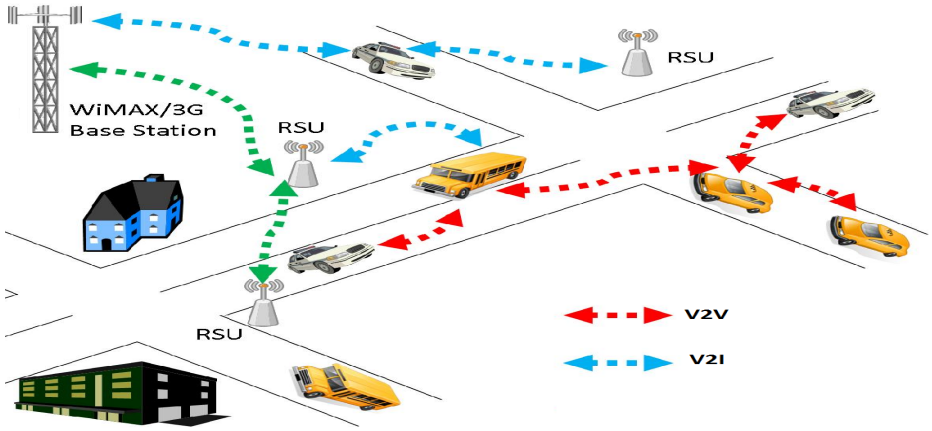

Figure 1.4 Vehicular ad hoc networks

The communication characteristics of a VANET are mostly based on a message called *BEACON*, periodically transmitted by each vehicle. Current VANET technology supports delivery of vehicle-to-vehicle BEACONs that are sent every 0.1 seconds. Therefore, networked cars can be extremely fast in warning their surroundings regarding events. By receiving BEACON messages, each vehicle therefore becomes 'aware' of what or who is around it, as well as its mobility characteristics. Accordingly, these messages will be the primary communication mean to acquire data for traffic monitoring.

The BEACON message contains a part that is fixed and carries time-stamped basic vehicle state information, such as senderID, position, direction, current speed, acceleration, with optional information also possible. BEACONs can be correlated with their senders via senderID. A vehicle in a VANET can continuously collect BEACONs from other vehicles along his path and from those, estimate traffic characteristics representing the evolution of traffic over time on the road network. Consequently, a huge amount of traffic condition data can be archived at a vehicle level. A vehicle stores vehicles' characteristics on each segment in information structures. Each structure consists of the following fields :

— SegmentID : The unique identifier of the road segment that this measurement belongs to.
— Time : Time of the measurement's creation.
— SenderID : The unique identifier of the vehicle that created this measurement.
— Position : Coordinates of the vehicle.

— Vinfo : Direction, speed and acceleration information of a vehicle.

After obtaining quantitative fine-grain traffic data, self-organized vehicular ad hoc networks can monitor the variation in traffic information of neighbour nodes in order to estimate real-time features.

1.1.3 Congestion

Congestion can be classified as recurrent and non-recurrent. Recurrent congestion refers to congestion that happens on a regular basis and usually occurs when a large number of vehicles use the limited capacity of the road network simultaneously. Non-recurrent congestion (NRC) in an urban network is mainly caused by incidents (accidents, vehicular breakdowns, police checks), workzones, special events (sport games, concerts, religious activities, political demonstrations), adverse weather[5].

Common causes of congestion are a widely investigated topic within the research community. Available research focus on studying the influence of a cause on traffic, such as the influence of weather, special events, incidents and workzones on traffic. In order to identify a set of variables representing spatial and temporal features of the components of congestion capable of distinguishing non-recurring congestion in an efficient way, we highlight from previous research key spatiotemporal characteristics and findings.

Inclement weather (rain, fog, snow, ice) has an impact on the fundamental macroscopic traffic flow variables (flow, speed and volume). It was also shown that microscopic traffic variables such as desired speed, desired acceleration and deceleration and minimum following distance parameters can be influenced during snowy road conditions for different reasons [6]. The free flow speed is defined by the speed driven when the driver is not influenced by nearby road users. Free speed is reduced to desired speed if the speed is influenced by other drivers, the road, characteristics of the vehicle, conditions such as weather and traffic rules (speed limits). For instance, drivers reduce their speed in order to avoid skidding in inclement weather. A reduction of desired speed, which reaches up to 30% for snowy roads, has been found. The reaction to adverse weather conditions varies between regions. As for the desired acceleration and deceleration, a slippery road reduces friction between road surface and tyres, thus drivers are not able to accelerate and decelerate as strongly as compared with dry road conditions, that is, maximum acceleration and deceleration decrease during snowy road conditions. Moreover, drivers reduce acceleration and deceleration to avoid skidding. Finally, drivers try to maintain a higher minimum following distance in order to cope with longer stopping distances caused by slippery roads [7]. From these findings, we assume that the features collected along a vehicle's trajectory and experience on other road segment that

can help infer weather conditions are : higher minimum following distance, reduced desired speed and higher travel times on some segments of the trajectory.

Regarding special events (sport games, concerts, religious activities, political demonstrations), previous researches have highlighted that they may lead people to travel towards the same destination in a very limited time interval, and then to leave the venue again in a very short time span [8]. The impact of a special event on traffic has been thoroughly studied in [9] for different demand categories of people going to the event (inbound traffic) and leaving after its end (outbound traffic). Special events may cause congestion depending on the intensity of ingress traffic demand or sharp traffic surge in concentrated time span. Thus a special event has an impact on the traffic behaviour in a specific region over time. Such an *impact region* can be defined as the list of congested segments of the road network around the special event. From these findings, we conjoncture that the most informative features along a vehicle's path that are the most informative of a special event traffic condition are firstly, the observed demand along the path in order to detect the presence of a sharp traffic surge. Also, if a vehicle experiences a NRC caused by a special event, then the vehicle is necessarily in the impact region of the event. Finally, if some road segments of the vehicle's path are inside the impact region, and the travel time on those segments are abnormally high, then we may associate this characteristic to a congestion caused by a special event.

Incidents (emergencies, accidents, vehicular breakdowns, road defects, police checks) and workzones also have an impact on the traffic flow variables. Particularly, these events physically block one or multiple lanes of the road segment. The blocked section of the road is called a problematic spot (Pspot) and can be defined by an interval of start and end point depending on the lane. No vehicles can traverse this section, thus reducing the capacity of the road. Trajectories of vehicles on a segment comprising a problematic spot cannot register coordinates of the blocked section. Regarding workzones, it was shown that if they happen at day time, especially at rush hours, their impact on traffic is severe, sometimes exceptionally larger than that of traffic accidents happening at the same time. On the other hand, if they happen at night, their impact is not that significant. Often times, workzones occupy the road segment a longer period of time than incidents because incidents are undesired and should be cleared as fast as possible. Thus, the duration of the Pspot is a good indicator of a workzone event.

Before we proceed, some definitions shall be highlited.

Definition 1 : A road segment is 5-tuple $r = \{s(x, y), e(x, y), l, r, TT_h\}$. Where s and e are the start and end point with location (x, y). l is the number of lanes. The road segment length is the Euclidian distance r between s and e. And the historical travel time TT_h

is a list of average expected travel times in seconds every five minutes interval along that road segment.

Definition 2 : A road network is a graph $RN = (N, E)$. Where E is the set of all road segments and N is the set of all road segments junctions.

Definition 3 : Trajectory is a sequence of connected road segments. It can be denoted as $T = (r_1, r_2, ..., r_n,)$ where $r_i \in E$.

Definition 4 : A trip is a vehicle movement from one place to another defined by the starting segment, the destination segment, and the departure time. A route is an expanded trip that means that a route definition contains not only the first and the last segment, but all segments the vehicle will pass along its trajectory.

With the above definitions, the traffic congestion problem can be formulated as follows. The observed travel time of a vehicle (oTT) along a road segment on an urban road network is composed of recurrent delay (D_{rec}) and non-recurrent delay, d, such as incident , workzone, weather or special event induced delays.

$$oTT = D_{rec} + D_{n-rec}$$

$$oTT = D_{rec} + \sum_{i=1}^{4}(x_i * d), where \tag{1.1}$$

$$\sum_{i=1}^{4} x_i = 1,$$
$$x_i = 0 or 1$$

D_{rec} is the expected recurring travel time TT_h that is location and time specific, based on the fact that the average speed on roads are usually similar at the same time of different days of the week. In transportation there is a big difference between weekday and weekend traffic. Weekday traffic is usually worse than the weekend. Expected values are derived offline using past historical data for each segment. The preloaded digital maps available on the vehicular nodes may provide this traffic statistic of the roads at different time of the day. If the observed travel time on the segment is higher than a threshold, which is determined as in [10] by multiplying the congestion factor c with the expected recurring delay, the travel time is said to be excessive.

$$\forall r \in E, oTT > (1 + c) * r.TT_h \Rightarrow oTT \text{ is excessive.} \tag{1.2}$$

Based on the road traffic condition, if congestion is detected on a specific segment and excessive travel time delay is observed, the oTT in an urban network is composed of a recurrent delay and a non-recurring delay due to the other components of congestion such as incident, workzone, bottleneck, weather or special event.

1.2 Problem definition

In the context of road traffic congestion, many studies have been devoted to highways rather than highly congested urban regions. In a highway scenario, the road section can be modeled as a network flow model that require flow conservation on all segment. The amount of flow entering an arc equals the amount of flow leaving the arc. In an urban scenario, on the other hand, each arc of the underlying graph has an associated positive gain or loss factor. Flow passing through the arc is magnified or diminished by a factor. Therefore, the design of accurate and scalable models for urban road networks is required which are intricate, complex networks and far more likely to be monitored by the traffic authorities.

Also, currently, traffic state cannot be directly measured everywhere on the traffic road network. Infrastructure operators monitor traffic on the network while using the least possible resources. The location where traffic flow needs to be improved varies highly and certainly, deploying highly sophisticated equipment to ensure the accurate estimation of traffic flows and timely detection of events everywhere on the road network is not feasible, due to the limitation in financial resources to support dense deployment and the maintenance of such equipment, in addition to their lack of flexibility. In fact, at this time, monitoring techniques do not scale to large cities. Infrastructure sensors that are in-roadway include inductive-loop detectors, which are sawcut into the pavement ; magnetometers, which may be placed underneath a paved roadway or bridge structure ; and over-roadway sensors such as video image processors that utilize cameras mounted on tall poles adjacent to the roadway or traffic signal mast arms over the roadway ; microwave radar, ultrasonic, and passive infrared sensors mounted in a similar manner are not capable of evolving over time and covering increasingly large geographic regions To scale to larger cities, advanced monitoring techniques should be deployed and must be capable of aggregating traffic data feeds at various levels of granularity.

Furthermore, the duration and timing of traffic events varies a lot. To evaluate the traffic state in real time, operators necessitate extensive data sources to guarantee the accurate evaluation of the traffic state. Current traffic data collection systems do not incorporate the ability of registring detailed information on the altering events happening on the road, such as vehicle crashes, adverse weather, etc. Operators require external data sources to retireve this information in real time. Besides, well-tailored data sources may not always be available for

a particular area of the traffic network. Future systems should enable continuous monitoring of the traffic condition along all roads of the traffic network based on real-time information.

Moreover, traffic bottlenecks are disruption of traffic on a roadway caused either due to road design, traffic lights, accidents, work zones, weather condition, special events occurring in the region, etc. There are two general types of bottlenecks, stationary and moving bottlenecks. Stationary bottlenecks are those that arise due to a disturbance that occurs due to a stationary situation like narrowing of a roadway, an accident. Moving bottlenecks on the other hand are those vehicles or vehicle behaviour that causes the disruption in the vehicles which are upstream of the vehicle. Moving bottlenecks can be caused by heavy trucks as they are slow moving vehicles with less acceleration and also may make lane changes. Also, moving bottlenecks can be active or inactive bottlenecks. If the reduced capacity caused due to a moving bottleneck is greater than the actual capacity downstream of the vehicle, then this bottleneck is said to be an active bottleneck. If the reduced capacity of the truck is less than the downstream capacity, then the truck becomes an inactive bottleneck. This portrayal shows that there are many causes to congestion or bottlenecks. Current methods only detect congestion but it's not enough, we should be able to better characterize the event causing it and distinguish between temporary induced traffic pattern change that is mitigated in a short period vs permanent pattern change. The existing methods can only quantify the spatial and temporal impact of the detected event. Agencies need to understand what is the cause affecting variability on their facilities and to what degree so that they can take the appropriate action to mitigate congestion.

The problem is that most existing works ignore the context information when proposing models in their study of traffic congestion. The context information can include : Time context (time of day), situation context (presence of traffic incident, workzone, weather condition or special event that occurred nearby in the region), trajectory context (travel time in the region), history context (past flows registered on road segments). While using all context dimensions will provide the most refined information and thus lead to the best performance, it is equally important to investigate which dimension or set of dimensions is the most informative if the task is to detect congestion, classify it or predict flow on a segment. The benefits of revealing the most relevant context dimension include reduced cost due to context information retrieval and transmission, reduced algorithmic and computation complexity and targeted active traffic control.

Many travellers have an overall sense of the status of traffic and the overall times until congestion at bottlenecks will likely start and end, based on their long-term experiences. People may be familiar with typical traffic patterns, some situations whether traffic states

of interest would be viewed as surprising. Current methods do not incorporate this overall sense, this experience, although the TIS is capable via its distributed systems to monitor and collect data from many sources. These data provide enough comprehensive information in order to better characterize the events detected. Current techniques fail to process the knowledge acquired from the data. With the widespread traditional traffic sensors and new emerging traffic sensor technologies, traffic data are exploding, and we have entered the era of big data transportation. In the big data era, techniques should be implemented to make use of the acquired information. Furthermore, at the current stage, the ITS is partially efficient since the vehicle is the only entity that is not contributing to the system. In fact, presently, vehicles are uninformative as they are not engaged in the process of traffic event detection. However, equipped with a communication technology, vehicles can exchange information and cooperate collectively so as to provide their input to the system because of the unpredictable nature of traffic and because of the myriad factors that affect traffic flows such as weather conditions, the behaviour of other drivers, traffic issues, and other events.

Knowing the flow of traffic heading toward a destination will give more insights about the expected demands in the near future. Indeed, traffic flow prediction allows advanced modelling so that traffic managers take early actions to control the traffic flow and prevent the congestion state. However, current models need to be improved so as to allow fast and more accurate prediction. In fact, in addition to traditional traffic sensors, a variety of data sources, such as lidar, radar, and video from surveillance cameras, have emerged in traffic flow prediction research [11]. The problem is that traffic flow prediction heavily depends on historical and real-time traffic data collected from various sensor sources, including inductive loops, radars, cameras, mobile Global Positioning System, crowd sourcing, social media, etc. As the data originate from different sources, their conversion is the most important step. In this process, the first obstacle is the amount of data collected which is increasing exponentially, and the second is its complexity. This makes data conversion increasingly difficult and highly time and resource consuming. Relevant data extraction and cleaning, as well as data reduction, is required. Each of these tasks has its own challenges, including defining what is relevant and what is noise, identifying one or the other, and extracting the useful data, given certain accuracy expectations. In sum, data aggregation poses many challenges when a variety of data sources are required in the process of data collection.

On the other hand, the problem with current traffic flow prediction models is their inadaptability of detecting and tracking the traffic patterns changes. There is a new pattern every time a non recurrent congestion occurs in the traffic flow and in this case, the model is not able to predict as accurately as when there is recurrent congestion. Existing approaches to traffic flow prediction do not adapt to the varying traffic situations because their distribu-

tion are memoryless, and they need a structure that will characterize the system at each step, not independently from the prior stage. To improve the flow prediction accuracy, the model should update from its normal path and track the changed traffic pattern, generating forecasts according to the new traffic pattern. A forecasting system will adjust its structure to accommodate the extraordinary patterns and the significant traffic flow pattern changes indicated will in turn drive the parameters of the short term forecasting system changing significantly to adapt to the traffic flow pattern change.

Finally, due to economic issues and lack of large scale deployment of connected vehicles technology, currently no datasets are available for researchers to test their models. Consequently, simulation is the main choice in the validation of models based on vehicular ad hoc networks. Although many discreet-event network simulators, such as ns-2, ns-3, OPNET, OMNet++, and QualNet, have been widely used by the researchers to validate their ideas and approaches, they cannot be used in ITS scenarios without an accurate vehicular mobility model because this may lead to unrealistic configurations of traffic distribution in the environment resulting in false representations of the network topology and network partitioning in isolated groups of nodes [6], [12], [13]. The insufficiency common to all these plateforms is that they ignore the spatiotemporal variability of mobility patterns (temporal aspect and external influence modules). The variability is due to both mobility rate of vehicles that changes periodically during the same day and daily during the week, and the characteristics of the environment that may dynamically change at any time due to external events that may occur, such as accidents, weather condition, special event and workzones.

These problems have led to the elaboration of the following research questions that we addressed thoroughly throughout the thesis :

— How are we going to collect microscopic and macroscopic traffic variables in real-time everywhere on the urban road traffic network since the location of congestion happening on the road varies highly and we don't know when and on what road segment traffic congestion is going to install ?

— Can vehicles estimate the cause of congestion they experience on the road and characterize it solely based on sensors they have on board and traffic flow theory and without any input from outside sources, such as weather information or police reports for incidents, etc. ?

— If vehicles can share their knowledge with the others in their vicinity, will they be able to better assess the traffic situation experienced or will they never conclude together on the real traffic condition because traffic is chaotic and multifaceted and each vehicle's trajectory experience is different from the other ?

— How can we predict flow on a target road segment with a model able to adapt to any varying traffic situations around the segment ?

— How are we going to test and validate any model we propose since connected vehicles technology is still at its early stage and with economic issues and lack of large scale deployment no dataset is available from any research ?

Thus, these questions led us to state the main research objectives presented in the next section.

1.3 Research objectives

This thesis aims at designing a traffic flow prediction framework that considers classification models able to estimate the cause of congestion via machine learning techniques. Mainly, by analyzing road traffic congestion through the advanced connected vehicles communication technologies, the goal is for vehicles to detect excessive congestion and classify collectively its cause on the segments of the trajectory in order to forecast flow on a target road segment so that infrastructure operators can take appropriate actions to mitigate traffic congestion. More specifically, the objectives of the thesis are :

— Collect measurable traffic features extracted by an advanced monitoring technology capable of aggregating microscopic and macroscopic traffic variables at various levels of granularity.

— Propose classification models based on the traffic features collected for inference on the cause of congestion and validate the models through the simulation of scenarios extended from a realistic urban city vehicular motion traces in order to build a synthetic dataset to feed the models for learning purposes.

— Implement a cooperation process to increase estimation accuracy and design data mining techniques for the real-time advanced distributed and continuous evaluation of road traffic condition. Test and validate the techniques via a combination of a microscopic urban mobility simulator, and a network simulator for the simulation of communication between connected vehicles.

— Design a traffic flow prediction framework taking into account historical flows as well as innovative features, such as real-time reports from connected vehicles and travel time along a trajectory for accurate forecasting of flow in an urban network in order to cope with the fact that existing approaches that do not adapt to the varying traffic situations. Validate the proposed framework through simulation generated scenarios

extended from a realistic data set of urban city vehicular motion traces.

1.4 Main contributions and their originality

The main originality of this thesis lies in the use of the next generation sensing technology of connected vehicles to identify road traffic events on the basis of exchanging traffic flow data between vehicles. This novel approach in this domain allow the real-time distributed detection and classification of the components of congestion in urban traffic. Moreover, if connected vehicles can detect congestion and cooperatively attribute a possible cause to it, we show that they can transfer this knowledge in real time to an entity able to accurately predict flow on a road segment. The traffic flow prediction framework we introduced aimed at evaluating anticipated traffic flow at future time frames on a target road segment based on real time feeds provided by connected vehicles and historical data.

The principal contributions consist of conceiving the models and framework, and can be described as follows :

— **Development of an algorithm for the distributed and real time advanced monitoring and evaluation of road traffic condition.** This first innovation resides in defining a set of qualitative and quantitative features that describe the real-time traffic state experienced by any vehicle along its trajectory via the connected vehicles technology. Since currently, monitoring techniques do not scale to large cities and traffic state cannot be directly measured everywhere on the traffic road network, infrastructure operators monitor traffic on the network while using the least possible resources. The location where traffic flow needs to be improved varies highly and certainly, deploying highly sophisticated equipment to ensure the accurate estimation of traffic flows and timely detection of events everywhere on the road network is not feasible, due to the limitation in financial resources to support dense deployment. Connected vehicles technology enables distributed and real time advanced monitoring.

— **Classification of congestion into its components via machine learning methods.** The methods we propose take the traffic features extracted by the connected vehicles into account for the inference. The deterministic classification tree, the naïve Bayesian classifier, the random forest and boosting techniques we presented are built on an understanding of the spatial and temporal causality measures of traffic data. This was the first work in the literature that try to estimate the cause of congestion based solely on traffic data extracted by vehicles. Conventional approaches use multiple external sources of information, merge them and compute an estimation of events

present on the road. Our contribution proved and validated that connected vehicles have enough technology on board, sensors and communication ability, to independently evaluate the cause of congestion. This contribution can assist transportation agencies in reducing urban congestion by developing effective congestion mitigation strategies knowing the root causes of congestion.

— **Proposing simulation generated scenarios to build a synthetic dataset to train the machine learning methods.** To learn representations via artificial intelligence, the methods require examples of inputs to train on, and each input is obtained by the vehicles of the connected vehicles. These examples are stored and constitute a dataset. The classification methods we propose need a dataset to estimate the cause of congestion. But due to economic issues and lack of large scale deployment of connected vehicles, currently no dataset is available to train the models. We relied on simulation to create a synthetic dataset. We used a microscopic urban mobility simulator, and a network simulator for the simulation of communication between CVs to create experiments on scenarios extended from a realistic urban city vehicular motion traces.

— **Implementation of a cooperation process to increase estimation accuracy.** We add an evaluation layer before fusion can take place on board of each vehicle in order to increase accuracy and lower false alarms that are comparable to security threats on the traffic network. We present distributed data mining techniques via connected vehicles to elaborate collectively a decision concerning the cause of traffic congestion on a road network. Our aim is improving the level of knowledge from exchanged messages to obtain deeper insight of traffic condition using decentralized cooperation between individual vehicles and local traffic behaviour. Observing the complex phenomena from the interactions between vehicles will allow more precise, efficient, and reliable view of the traffic condition. The originality of the contribution is the generation of a dataset for association rules mining to extract more knowledge and implementation of the rules on board of the vehicles for analysis and evaluation of the cause of urban traffic congestion. We evaluate a Voting Procedure, Belief Functions and a Data Association Technique.

— **Forecasting of short term traffic flow on a target road segment via input from connected vehicles.** The prediction of traffic flow constitutes a fundamental contribution of this thesis. The basis of this prediction lies in the fact that we integrate the impact of various events into the forecasting of traffic flow on a target road segment. We propose a Deep Neural Networks (DNNs), and tackle the problem by learning the target DNN in a multitask learning technique. We conjecture that when the tasks

involved in MTL are semantically connected, a larger improvement in predication accuracy can be obtained. More specifically, MTL can be more effective when we can encode the instances from different tasks using the same representation layer expressing similar semantics. Using historical flows and well engineered features, such as real-time reports from connected vehicles and travel time along a trajectory for accurate forecasting of flow in an urban network, the model learns a representation that takes into account the various events that vehicles realistically encounter on the segments along their trajectory. They may come across incidents, workzone, inclement weather, special events or recurrent congestion. All these situations are assessed by the connected vehicles and are modeled by creative features to be fed to the DDN for the sake of learning to predict traffic flow.

In general, traffic flow prediction allows advanced modelling because knowing the volume of traffic heading toward a destination will give more insights about the expected demands in the near future. The proposed techniques for data collection and classification and the developed framework for flow prediction will help infrastructure authorities improve the network traffic flow and thus reduce traffic congestion.

1.5 Thesis structure

Having defined the basic concepts related to the work in this thesis, described the research problem, cited the objectives of the thesis and briefly explained our major contributions, chapter 2 will review the literature related to each element of the research problem that we described. An analysis of the limitations of existing work and the gaps that must be filled will also be elaborated throughout this chapter. In chapter 3, a detailed description of our research work and published articles is given, and the relationship between our objectives is emphasized.

Chapter 4 presents the full text of the article titled "Distributed Classification of Urban Congestion Using VANET", which was published in *IEEE Transactions on Intelligent Transportation Systems*. The main contribution of this article lies in the evaluation of machine learning methods for the classification of congestion into its components taking traffic features collected from connected vehicles into account for the inference on the cause of the non recurrent traffic congestion. Because detecting congestion is not enough, this article proposes that congestion can be further classified as recurrent and non-recurrent congestion. In particular, NRC in an urban network is mainly caused by incidents, workzones, special events and adverse weather. The framework considers the real-time distributed classification

of congestion into its components on a heterogeneous urban road network using connected vehicles. Models are built on an understanding of the spatial and temporal causality measures and trained on synthetic data extended from a real case study of Cologne. The performance evaluation shows an estimation accuracy of 87.63% for the deterministic Classification Tree (CT), 88.83% for the Naive Bayesian classifier (NB), 89.51% for Random Forest (RF) and 89.17% for the boosting technique. This framework can assist transportation agencies in reducing urban congestion by developing effective congestion mitigation strategies knowing the root causes of congestion.

Chapter 5 presents the full text of the article titled "Cooperative Evaluation of the Cause of Urban Traffic Congestion via Connected Vehicles ", which was submitted in *IEEE Transactions on Intelligent Transportation Systems*. In this article, firstly, the framework for the classification of congestion previously developed, where each vehicle can detect individually excessive congestion and attribute a cause to it, was implemented on board of each vehicle. To obtain deeper real-time insights of traffic conditions and improve estimation accuracy, in the article, methods using decentralized cooperation between individual vehicles where applied. A distributed data mining based methodology to elaborate a decision concerning the cause of traffic congestion on a road network via emerging connected vehicle technologies was developed. The performance evaluation shows that the proposed methods enhance the estimation of the cause of congestion, reduce detection time and trigger less false alarms. This implies that the complex traffic phenomena is better observed through the interactions between vehicles exchanging messages.

Chapter 6 presents the full text of the article titled " Prediction of Traffic Flow via Connected Vehicles", which was submitted in *IEEE Transactions on Intelligent Transportation Systems*. The article addresses the problem of traffic flow prediction. It integrates the fact that vehicles traveling along a trajectory can detect excessive congestion and collectively attribute a cause to it into the forecasting of traffic flow on a target road segment. A multitask learning deep neural network technique is proposed and takes as input the real-time reports from connected vehicles as well as other relevant features, such as travel time, historical flows. The DNN input features take into account both macroscopic and microscopic traffic variables in the prediction of traffic flow. The results show our approach significantly outperforms existing approaches that do not adapt to the varying traffic situations. DNN learned historical similarities between road segments, in contrast to using direct historical trends in the measure itself, since sometimes trends may not exist in the measure but do in the similarities.

Chapter 7 presents a general analysis and discussion regarding the strong points and limitations of our research work in this thesis. Finally, chapter 8 concludes the thesis by presenting

a summary of our work and a discussion about future potential research avenues that could extend our work.

CHAPTER 2 LITERATURE REVIEW

In this chapter, we discuss the recent work that has been done in the research areas related to this thesis. First, the work on congestion detection will be examined. This includes the state-of-the-art methods based on infrastructure and those based on the vehicles. Then, a description of the research related to the classification of congestion will be elaborated. This includes the point of view from the transportation engineering which is an offline approach. We also present the studies done online which are in traffic event analysis and anomaly detection systems. Afterwards, current approaches for the evaluation of the cause of congestion by vehicular ad hoc networks will be reviewed. Finally, the literature related to the prediction of traffic flow will be investigated from two perspectives : the type of traffic data source used to collect the data and the technique used to model traffic data as they are factors that affect the forecasting accuracy. At the end of the chapter, we carry out a thorough analysis of the described work to show its limitations and identify the research gaps that need to be filled.

2.1 Detection of congestion

In this section, we review the works done on congestion detection. In transportation engineering, they rely mainly on the methods based on the infrastructure such as induction loops, cameras or sensors in order to collect traffic data and detect congestion. On the other hand, other methods to detect congestion on a road segment are based on collection methods done by the vehicles, such as floating car data and recently using connected vehicles technologies. We mainly focus on the connected vehicles' ability to detect congestion. Particularly, we present the studies in this field that attempt to detect traffic congestion by vehicle to Road Side Units (RSU) communications. We also present the studies that avoid the installation of RSU, and detect congestion via V2V communications.

2.1.1 Methods based on the infrastructure

In transportation engineering, they rely mainly on the methods based on the infrastructure to collect data and detect road traffic congestion [3]. While single inductive-loop detectors give direct information concerning vehicle passage and presence, other traffic flow parameters such as density and speed must be inferred from algorithms that interpret or analyze the measured data. When these parameters are calculated from inductive-loop data, the values do not have sufficient accuracy for calculation of link travel time [14]. Evaluation of over-roadway sensors

show that they provide an alternative to inductive-loop detectors. Particularly, video image processing automatically analyze the scene of interest and extract information for traffic surveillance and management [11]. Although the traffic flow parameters measured with over-roadway sensors satisfy the accuracy requirements of many applications, infrastructure-based detectors provide fixed-points and short-section traffic information that is extracted from vehicles passing through the detection zone only [1].

2.1.2 Methods based on the vehicles

We present here the methods to detect congestion on a road segment that are based on the vehicle, such as floating car data of probe vehicles, and recently using connected vehicles technologies.

Probe vehicles or mobile sensors appeared as a complementary solution to fixed sensors for increasing coverage areas and accuracy without requiring expensive infrastructure investment. In [15], they detect congestion with data collected from probe vehicles. In [16], they assess the traffic variables from cellular feeds and identify the traffic condition based on speed of vehicles.

On the other hand research in VANET proved and validated that via distributed computing and mobile communication, vehicular networks can efficiently assess the status of the road. Based on the road traffic condition, the congestion status can then be identified. Vehicles exchange via wireless communication microscopic or macroscopic traffic variables and different studies apply different schemes for the detection of congestion. In some studies they attempt to detect traffic congestion by vehicle to Road Side Units communications [17]. While in others, to avoid the installation of RSU, vehicles exchange the traffic data by vehicle to vehicle communication [18]. Once the traffic data is gathered, the schemes introduced in the literature using VANET mainly use characteristics such as traffic density, traffic speed, traffic volume or estimated traveling time to detect congestion.

In [19], they defined a saturated traffic density (SDi) of each road segment i. The density level is the ratio between the number of vehicles at each road segment and the road segment length per lane on the respective road segment. The congestion level is correlated with the traffic density at each road and is set as low congestion level of 0.025 to a high congestion level of 0.125 a vehicle per meter in each lane. In [20], they detect urban traffic congestion with single vehicle and distinguish congested from free-flow road segments. In [21], they integrated feature-level and decision-level information fusion to improve the reliability of congestion detection. Lots of works utilize machine learning mechanisms to classify the traffic state into congested or free-flow [22], [23]. To classify the level of congestion, [24] proposes

a traffic congestion quantification process based on fuzzy theory. The fuzzy-based detection mechanism takes the traffic density estimate and the vehicle's speed as input parameters, and provides the traffic congestion level as output parameter (free-flow, slight congestion, moderate congestion, severe congestion). Similarly, estimation of congestion degree is done in [18], where LoC, the level of congestion has values from 0 to 1, ranging from free flow to severely congested. Moreover, traffic congestion notification combines vehicular cooperation and human assistance to determine the overall congestion degree and the area of congestion. Research on VANET proved and validated that via distributed computing and mobile communication, vehicular networks can efficiently assess traffic conditions on roads. These approaches only detect congestion and do not clarify if the observed congestion is due to recurrent or non recurrent congestion. They cannot be used to classify the congestion into its components.

2.2 Classification of congestion

In this section, we present the studies that tackled the classification of the cause of traffic congestion. In the field of transportation in civil engineering, understanding how much of the total congestion is due to NRC has been thoroughly studied for both highway [25] and urban traffic [5] in an offline approach. From the networking point of view, research has not yet tackled the problem of classifying the congestion detected into its components. Instead, in this field, they carry out traffic event analysis and present anomaly detection systems which are online approaches. We will present studies in each of these domains and explain how they are not solving the problem of classification of congestion into its components.

2.2.1 Offline approach

In [25], the congestion pie chart was introduced to visualize the contribution of the components of congestion. Data requirements for the studies include traffic incident logs from agencies, data on travel time along routes over days for a specific time of day, weather, workzones and special events data during the time interval. For instance, large volumes of information at very high spatial and temporal resolutions from different sources are required for the methods to be applied. Moreover, research on detecting NRC events has only recently gained importance. In the work [10], they propose methods for NRC event detection on heterogeneous urban road networks based on link journey time (LJT) estimates. The LJT data is estimated from vehicle journey times that are obtained by matching the data of automatic plate number recognition cameras. Since an NRC would cause an unexpected delay with respect to expected travel times, they detect statistically significant clusters of high LJTs.

Expected LJTs capture the recurrent nature of traffic. Excess LJT is the difference between the observed and expected LJT. If the observed LJT is higher than a threshold, which is determined by multiplying the congestion factor c with the expected LJT, the LJT is said to be excessive. The data quality of the LJT estimates depends on the size of the sample used to compute it. The higher the number of vehicles' journey times that are used to estimate an LJT, the higher the data quality of the estimated LJT is. They suggest important values regarding the threshold to determine whether or not an LJT is excessive, high confidence episodes and localization index in order to detect reasonable NRC. These methods not only need extensive datasets, but they are also not deployed in real-time. In real-time, valuable information with regard to the potential impacts of the detected NRC can be disseminated to individual drivers and traffic management centers so that appropriate proactive strategies for recovering traffic conditions back to normality can be set in place.

2.2.2 Online approach

2.2.2.1 Traffic event analysis

On the other hand, the effect of events on traffic has been studied in the field of data mining [26]. For example, the impact of incidents on traffic is studied in [27]. Congestion caused by special events is examined in [8]. The impact of inclement weather on the fundamental traffic flow variables is investigated in [28]. Workzone events create conditions that are different from both normal operating conditions and incident conditions and they are investigated in [29]. There are also several studies that mainly concern the cause of the events, aiming at how to design the network or re-direct the traffic flows to avoid the delay of events [30]. Knowing the event, the studies analyse its impact on the traffic variables. We aim at doing the inverse ; we need the traffic variables that can help infer the component causing the excessive delay.

In [31], a software agent based approach to disseminate critical information during critical situations is proposed. The cognitive model which performs pull and push processes assumes that an agent platform exists in vehicles, base stations and regional transport station. The proposed agent based information dissemination model is a network and application architecture. In [32], the authors put forward an event-driven architecture (EDA) as a mechanism to get insight into VANET messages to detect different levels of traffic jams ; the mechanism takes into account environmental data that come from data external sources, such as weather conditions, web services or onboard sensors. EDA is a software architecture where relevant real-world activities are reflected as events in the lowest layers of the EDA architecture. This method is not local and self-organized to generate events based on the real-time relevant information extraction because it needs external data sources for inference.

Context mining is complementary to network-sensing technologies, when properly aligned in space and time, they become essential to understand transport-related phenomena [33]. Context-awareness is the potential to access available semantic information such as time, location, weather, temporary events and other attributes. The context information used in [32] fuses different data-sources (internal sensors, external data-sources from Internet or cloud services and passenger sensors such as smartphone) in almost real-time. In [34], the authors propose a traffic condition detection algorithm in V2V communication. They suppose that context can either be derived from the activity of the individual cars' electronic helpers like ESP (Electronic Stability Program) or ABS (Anti-locking Brake System), or alternatively, sensors embedded in the individual vehicle may provide this information. In this research, we do not use those indicators to detect a weather condition, instead, we use traffic flow theory because it was shown that a weather condition such as (rainy, snowy, slippery, foggy) has an impact on the fundamental traffic flow variables, flow, speed and density [28]. A detection based on a theory has a stronger foundation than that based on the hypothetical assumptions of car sensor integration rate. Vehicles need to be context aware and able to consider multiple but adequate explanatory sources, well-tailored information won't always be available, particularly in dynamic urban networks.

2.2.2.2 Anomaly detection systems

Recent studies have attempted to develop algorithms to detect traffic anomalies, or outliers also known as traffic patterns that do not conform to expected behaviour [35]. The majority of the studies on abnormal traffic pattern using macroscopic and microscopic traffic variables attempt to detect anomalies prior to the occurrence of an incident [36]. The aim for the analysis of the traffic variables in this case is to identify incident precursor phases. In [37], the authors propose a system for automatic detection of problematic road conditions. They analyze possible events on some lane by using vehicle's lane-changing information and it is designed to enhance the existing incidents detection techniques. Beside the lane-changing information, [38] makes use of more information to derive incidents, in sensor-assisted VANET. The scheme extracts the distribution of vehicle footprints on the road, the lane-switching patterns and the traffic density. Vehicles record their *footprints* (i.e., the geographical position at each sampling time point) periodically, and the footprints are aggregated and analyzed to infer the occurrence and locations of problematic conditions.

While traffic incidents can lead to congestion and understanding the precursor phases by analyzing the traffic variables is a first step, workzones, weather and special events are also events that disrupt the normal flow of traffic. The fact that incidents are not the only cause

of NRC limit the usefulness of existing automatic incident detection methods for identifying NRC on urban road networks. These methods cannot solve the classification of congestion into its components problem.

2.3 Evaluation of the cause of congestion

Since urban traffic is essentially unstable, chaotic, and unpredictable, each vehicle assessment is not enough, an evaluation process is essential in order to elaborate a decision on the cause of congestion. This would result in a more precise, efficient, and reliable view of the traffic condition by observing the complex phenomena from the interactions between vehicles. In this section, we present the studies in the literature that try to solve this problem, and they mainly use data fusion as a technique to evaluate traffic events.

Data fusion algorithms take the data collected by the connected vehicles and use it to improve : the reliability of a judgment by the contribution of redundant information ; or the interpretation ability by the provision of complementary information. Particularly, a large portion of literature has been proposed for the distributed data fusion for uncertain reasoning in ad hoc and dynamical networks [39] [40] [41] [42]. In [39], they introduced belief functions to combine and fuse data in vehicle for the management of uncertainties about events in vehicular networks. The theory of belief functions is a generalization of the Bayesian probability theory. Belief functions combine degrees of confidence about events reported in exchanged messages. Their methods were tested and compared using a Matlab simulator where roads are divided into segments and one event is considered per segment. Exchanged messages via V2V inform of the presence or the absence of events. Their methods do not learn from the data they collect and fuse. Also, it should also be remarked that V2V algorithms for combining and fusing data are very different from algorithms developed in Vehicle to Infrastructure (V2I) communication applications [43] [4]. In the latter study, a centralized module combines collected data and disseminates global information.

Specifically, concerning spatio-temporal events such as traffic jams, some studies propose methods that tackle both the information dissemination and information fusion problem. In [40], belief regarding the presence of an event on a geographical point is obtained by : discounting [41] neighbouring information according to their distance from the point ; then combining the obtained information [42]. In [40], authors propose to use the cautious combination rule [44] to fuse information located on a same road segment. In [45], they presented a system to manage information about uncertain events, but unlike the model in [40], was the choice of the event dissemination strategy considered. Each vehicle sends new events or repeats received one. A choice has been undertaken to keep combinations of messages in each vehicle and to

not diffuse it, each driver making its own overview of the situation, the environment being not overloaded with partial fused messages. With the use of belief functions, an overview of the situation regarding each event can be given to the driver such that each event is associated with a degree of confidence and then broadcasted to the outside world if necessary. Also, the model allowed events of the same type to be present on the same road segment, for instance : different accidents, different parking space, etc. It implied the necessity to assure a procedure to determine identical events. For validation, the model was implemented and tested using Hong-Ta Corporation (HTC), an application using smartphones. The application proposed required driver assistance to send the events and the authors proposed that camera or sensors might be installed in vehicles in such a way to automatically detect events.

In [41], a method is proposed to exchange and manage information about events on the road in V2V communication taking into account non-spatial events and spatial extent of a traffic jam. The performance of the method is measured by considering the adequacy between the information given to the drivers in each vehicle and the reality. Authors propose to divide map traffic lanes into small rectangular areas named cells. The map is composed of horizontal and vertical two-way streets in particular a traffic lane is composed of $NbSimCells$ cells depending on the type of event. Authors suggest that sensors might be used to detect events in order to create messages automatically, without driver assistance.

In [42], different strategies are compared to manage as best as possible fusion of acquired information, message aging of local events and dynamics and spatiality of traffic jams. They extend the work in [41] by developing new methods based on the notion of update and by proposing a way to automatically compute the message aging (by discounting or reinforcing) using historical data. Their work distinguishes local events (such as accident) and spatial events (such as traffic jam). The list of sources is kept in vehicle database in order to consider finely the dependence between messages, and use the most suitable combination operator (either the conjunctive rule or the cautious rule) to combine information.

In [42], since their influence mechanism predicts the transfer of the traffic, different causes of congestion require different types of management for the mechanism to work properly and not generate false influences. They stated that unlike traffic jams, the spatiality of fog blankets does not depend on maps and to manage this spatial event, roads are divided into cells, without taking into account traffic directions. In other words, if a fog blankets event is present on one side of a traffic lane, it is also certainly present on the opposite side. The influences of a fog blankets event concern surrounding cells, without any certainty of its presence or its absence. None of the aforementioned methods specify any procedure to gather data for the sake of learning. In fact, data mining techniques such as clustering, association,

classification, have been applied in VANET to extract useful patterns and information [26].

2.4 Prediction of traffic Flow

The traffic flow prediction problem aims at evaluating anticipated traffic flow at future time frames on a target road segment. In this section, we present the literature related to the prediction of traffic flow. We review it from two perspectives : the type of traffic data source used to collect the data and the technique used to model traffic as they are factors that affect the forecasting accuracy. Traffic flow prediction techniques can be mainly classified in three categories : 1) parametric approach ; 2) nonparametric approach ; and 3) hybrid approach.

2.4.1 Parametric approach

The main techniques used in this category are time-series models, AutoRegressive Integrated Moving Average (ARIMA)-based models [46] and Kalman filtering [47]. In [48], they applied an ARIMA model for traffic volume prediction in urban arterial roads. Many variants of ARIMA were proposed to improve prediction accuracy, such as Kohonen-ARIMA (KARIMA) [49], ARIMA with explanatory variables (ARIMAX) [50], vector autoregressive moving average (ARMA) and space–time ARIMA [51], and seasonal ARIMA (SARIMA) [52]. Other types of time-series models were also used for traffic flow prediction such as the statistical models. They make the assumption of stationarity of the underlying process. This assumption is often violated as observable traffic conditions can evolve differently at different times. Also, the linearity of the time series approach presents an inconvenience for traffic prediction. Traffic flow has stochastic and nonlinear nature, unfortunately, even an enhanced ARIMA cannot accurately predict flow in the presence of accidents. ARIMA, due to its delayed reaction, is not an ideal method to use in the case of events which cause sudden changes in the time series data. If we know per say, from police event streams, that there is an accident (say, 30 minutes) ahead of us, we may be able to predict its delays and account for it. On the other hand, historical data can be used to identify similar accidents, i.e., with similar severity, similar location and during the similar time, so that we can use their impact on average speed changes and backlog to predict the behaviour of the accident in front of us. For example, an accident that may happen between 4 :00PM and 8 :00PM on a particular road segment might cause 5.5 miles of average backlog ahead of the accident location. If the same accident happens between 8 :00PM and midnight the backlog will be 2.5 miles. In addition, these techniques predict traffic flow on each road segment separately. Since transportation networks are complex and much correlated, it is crucial to predict traffic flow from a network perspective. Moreover, while time-series analysis models are probabilistic, they are ignorant

of the underlying process that generates the data. Thus, time-series-based approaches are more prone to large errors in traffic flow forecasting.

2.4.2 Nonparametric approach

Nonparametric regression [53] is a widely used technique. In [54], an online boosting regression technique that ensures traffic prediction under abnormal traffic conditions was proposed. Otherwise, boosting is disabled. In [55], a support vector regression was used to establish the prediction model, whereas particle swarm optimization was used to optimize the model's parameters. Among all of these techniques, neural-network-based forecasting had the best performance in terms of prediction accuracy and are considered to be relatively effective methods because of their well established models.

2.4.2.1 Neural Networks

A panoply of artificial neural networks (ANNs) were proposed to predict traffic flow [56]. Typical computational intelligence-based forecasting methods mainly include the back propagation (BP) neural network [53], radial basis function (RBF) neural network [57], recurrent neural network [58], time-delayed neural network [59], and resource allocated networks [60]. Particularly, deep learning is a neural network of more than one hidden layer. This technique has attracted researchers from various domains as it considers complex correlations between features and outputs. Besides the factors of scope, data resolution and technique used to model the traffic, we will compare the works in this approach with regards to the features used to train the models and the type of traffic data source used to collect the data.

In [61], they propose a stacked auto-encoder model to learn generic traffic flow features by considering the spatial and temporal correlations. The model is trained in a greedy layer-wise fashion. The traffic data are collected every 30s from over 15 000 individual detectors, which are deployed state wide in freeway systems across California. Their input features consists of traffic flow data at previous time intervals, on the target road segment. The target road segment is the link of interest in the road network where the model wants to predict flow on. Considering the temporal relationship of traffic, to predict the traffic flow at time interval t, they use the traffic flow data at previous time intervals. In this study, their simulations indicate that four past time intervals of 15 minutes are enough to get good performance.

On another hand, recent work has shown that it is possible to jointly train a general system for solving different tasks simultaneously [62], multitask learning (MTL). If the tasks can share what they learn, the learner may find it is easier to learn them together than in isolation.

MTL is one way of achieving inductive transfer between tasks. The goal of inductive transfer is to leverage additional sources of information to improve the performance of learning on the main task.

In [63] and [62], they train a MTL model to predict flows on links. Unlike traditional traffic flow forecasting that predicts a future flow of a certain link only using the historical data on the same link, which is also called single-link traffic flow forecasting, the authors propose multilink forecasting models, which take the relations between adjacent links into account. Single-link forecasting approaches ignore the relationships between the measured link and its adjacent links. In fact, each link is closely related to other links in the whole transportation system. The multilink model predicts traffic flows using historical traffic flow data from all of the adjacent links. The features in [63] are flow data collected from sensors on the road. In [64], they propose a combination of multitask learning and an ensemble learning method bagging, for traffic flow forecasting. In [62], they propose a deep architecture that consists of two parts, i.e., a deep belief network (DBN) at the bottom and a multitask regression layer at the top. In a transportation system, all roads and entrance–exit stations are connected to each other. There is a lot of shared information among these roads and stations. The data are collected from inductive loops continuously collecting data in real time for more than 8100 freeway locations throughout the State of California. In [65], they proposed approaches showing significant improvements in prediction accuracy when compared to baseline predictors but their focus lies on highways that are one-directional road segments, whereby usually in the inner cities the impact of traffic is a multi-dimensional problem, evolving in a 2D, more complex route network.

Current research on traffic flow prediction mainly focuses on data traffic history and neglects other conditions affecting traffic. In [66], they investigate and quantify the impact of weather on traffic prediction in a freeway scenario. They admit that transportation systems might be heavily affected by factors such as accidents and weather. But they just considered the weather factor. They claim that inclement weather conditions may have a drastic impact on travel time and traffic flow. Their MTL architecture incorporate deep belief networks for traffic and weather prediction and decision-level data fusion scheme to enhance prediction accuracy using weather conditions. The traffic flow predictions provided by their approach use past values of the traffic flow and the current weather data is fused to provide future traffic flow prediction. They state that their scheme avoids compounding prediction errors that may ensue had weather data been predicted rather than been used as real information. Traffic flow is measured every 30 s using inductive loop sensors deployed throughout the freeways.

2.4.3 Hybrid approach

Some studies have investigated hybrid approaches [67]. To obtain adaptive models, some works explore hybrid methods by combining several techniques. Although the aforementioned hybrid models are flexible, they do not fully take profit from spatial information collected from the whole road network. Moreover, these studies rely only on information collected by sensors such as the Global Positioning System, loop detectors, and smart-phones. In traffic event analysis, the effect of events on traffic prediction has also been studied in the fields of data mining and transportation engineering. The majority of these studies focused on real time event/outlier detection using probabilistic or rule-based approaches (e.g.,[68], [69], [35]). There are also several studies that mainly concern the cause of the events, aiming at how to design the network or re-direct the traffic flows to avoid the delay of events [35]. However, none of these studies incorporate events into traffic flow prediction techniques, and hence fail to provide realistic forecasting in the presence of events.

2.5 Analysis and limitations

As the complexity of traffic increases, data collection methods that are based on the infrastructure, such as the loop detectors and those that rely on floating car data are not scalable for the urban network because data collection efforts are too expensive to replicate on a large scale or on a continuing basis. Due to real-time constraints much more information extraction techniques are needed to extract transport-relevant parameters. New technology such as connected vehicles can be used to improve the accuracy, timeliness, and cost efficiency of data collection. Real-time distributed monitoring refers to the process by which macroscopic and microscopic traffic variables are collected by vehicles themselves. Such systems are still at an early stage and their development has been hampered by the inability of existing monitoring systems to deliver traffic flow data with sufficient spatial granularity and timeliness. The larger coverage due to the distribution and high mobility of connected vehicles is taken as an advantage for urban detection of congestion.

The studies presented in this review on connected vehicles validated that via distributed computing and mobile communication, vehicular networks can efficiently detect road traffic congestion. However, the studies proposed only detect congestion and do not clarify if the observed congestion is due to recurrent or non-recurrent congestion. Furthermore, they cannot be used to classify the congestion into its components. The problem is that the research in this field is not evolving at the same pace as that in the transportation engineering area.

In order to classify the cause of congestion, data requirements for the studies done in civil

engineering in the transportation field include traffic incident logs from agencies, data on travel time along routes over days for a specific time of day, weather, workzones and special events data during the time interval. For instance, large volumes of information at very high spatial and temporal resolutions from different sources are required for the methods to be applied. Also, they not only need extensive data sources, but they are deployed offline. In real-time, traffic management centers are interested in detecting NRC and specially its cause so that they can set in place appropriate proactive strategies for recovering traffic conditions back to normality. Online approaches proposed in the literature for traffic event analysis in the connected networks field also require extensive data sources. The context information fuses different data sources (internal sensors, external data-sources from Internet or cloud services and passenger sensors such as smartphone) in almost real-time. Although vehicles need to be context aware, we see that well-tailored information won't always be available, particularly in dynamic urban networks. Due to real-time constraints much more information extraction techniques are needed to extract transport-relevant parameters. An automatic method where contextual information is taken into account in the assessment process is needed.

On the other hand, if each vehicle classifies individually the cause of congestion, the assessment and classification is done locally at a vehicle level. If one vehicle sends a false alarms, it spreads uncertainty among vehicles and this in turn causes more congestion. Sending false information disrupts the proper network operation and presents a threat to the traffic network. This makes exploring the cause of congestion at a vehicle level a partial solution. The studies reviewed in this thesis for the evaluation of the cause of congestion focus on data fusion techniques. Because traffic is multifaceted, we warn that a vehicle by itself has partial knowledge about the road condition, it knows to some degree the traffic condition surrounding it, so fusion of data alone is a limited solution. If we collect the fusion results for the sake of learning, we believe that knowledge can be acquired from the fused data that the vehicles exchange in a presence of a particular road condition. In urban networks, vehicles are repeatedly faced with situations where they encounter congestion. Vehicles will have to repeatedly determine its cause based on the variables they collected. They should be intelligent enough to learn from their experiences. Currently, there is no mechanism that is able to extract valuable knowledge from the situations experienced by the vehicles. Every situation should be a suite of instances learned for better decision making because the monitoring currently done by the proposed schemes does not allow for summarizing valuable knowledge. Also, to improve the accuracy of the evaluation of the cause of congestion, vehicles should elaborate a decision collectively.

Finally, a prediction algorithm uses real-time feeds and applies an advanced modeling technique combined with historical data to predict the future traffic flow on a segment. To the

best of our knowledge, no research was done on the use of connected vehicles as real-time feeds to the prediction model. In fact, in all studies in the literature, data is collected from fixed monitoring equipments, such as sensors and CCTV cameras, or using mobile data sources such as floating GPS data and SMS, social data feeds [70]. The type of traffic data source is a very important factor as it affects the efficiency and accuracy of the prediction algorithm. Also, most of the studies have been devoted to highways rather than highly congested urban arterials, which are far more likely to be monitored by the traffic authorities. Therefore, the design of accurate and scalable traffic prediction tools for urban road networks is required.

The majority of the techniques in the parametric approach focus on predicting traffic in typical scenarios (e.g., morning rush hours), and more recently in the presence of accidents or a weather condition. Existing techniques are only applicable to predict one of the scenarios. ARIMA prediction model is more effective in predicting the speed in normal conditions but at the edges of the rush-hour time (i.e., the beginning and the end of rush hour), the HAM model is more useful. This becomes even more challenging when considering different causes for congestion, e.g., recurring (e.g., daily rush hours), occasional(e.g., weather conditions), unpredictable (e.g., accidents), and temporarily—for short-term (e.g., a basketball game) or long-term (e.g., road construction) congestions. The techniques consider traffic flow as a simple time-series data and ignore phenomenons that particularly happen to traffic data. For example, for generic time-series, the observations made in the immediate past are usually a good indication of the short-term future. However, for traffic time-series, this is not true at the edges of the rush hours, due to sudden speed changes. The statistical approaches, by their very nature the mathematics of collecting, organizing, and interpreting numerical data, can provide more insights on the mechanisms creating and processing the data. On the other hand, the statistical approaches frequently fail when dealing with complex and highly nonlinear data and suffer from the curse of dimensionality. Finally abnormal traffic patterns caused by non recurrent congestion may deteriorate the performance of these models [71]. Nevertheless, under most situations, extreme values are of primary interest in forecasting the change in traffic conditions.

In general, literature shows promising results when using neural networks models as they are used as benchmarking methods for short-term traffic prediction. But since simulations have shown that one hidden layer would not be enough to describe the complicated relationship between the inputs and the outputs of the prediction model, deep learning has attracted researchers in this domain as it considers complex correlations between features and outputs. But even the studies that trained a deep neural network to solve different tasks simultaneously, used MTL with flows only as inputs to the model. No studies integrate the impact of various events into the forecasting models by considering spatiotemporal characteristics

of traffic in training of the models. In fact, no studies have tackled the problem of training an MTL by analyzing the tight correlation between traffic data and external factors for the prediction of traffic flow in an urban traffic network. It should be noted that the prediction of traffic flow under atypical conditions is evidently more challenging than doing so under typical conditions and, hence, much desired by operational agencies.

CHAPTER 3 METHODOLOGY

The aim of this thesis is to analyse road traffic flow in the transportation system through the connected vehicles technologies in order to mitigate congestion. To this end, techniques for data collection via connected vehicles technology, cooperative classification of the cause of congestion and a framework for flow prediction were envisioned as announced in section 1.3. To attain these objectives, the work was carried out in three main phases. This chapter aims at explaining the different steps that led to the realization of the objectives and connecting them to the work presented in the subsequent chapters.

3.1 Phase 1 : Classification of traffic congestion

In the first phase of our work in this thesis, our efforts were directed towards the achievement of the first two objectives. The work consisted of : 1) first, collecting measurable traffic features extracted by an advanced monitoring technology capable of aggregating microscopic and macroscopic traffic variables at various levels of granularity, and 2) second, proposing classification models based on the traffic features collected for inference on the cause of congestion. These objectives were attained in the first article titled "Distributed Classification of Urban Congestion Using VANET" presented in chapter 4.

3.1.1 Components of congestion

Firstly, vehicles should be able to detect congestion via the connected vehicles technology while travelling along routes of their trajectory. We use the travel time variable to assess the traffic condition as it is the most accurate metric that CVs can calculate and that is what distinguishes them from conventional data collection methods. In fact, the observed travel time of a vehicle (oTT) along a road segment on an urban road network is composed of recurrent delay (D_{rec}) and non-recurrent delay, d, such as incident , workzone, weather or special event induced delays.

$$oTT = D_{rec} + D_{n-rec}$$

$$oTT = D_{rec} + \sum_{i=1}^{4}(x_i * d), where \tag{3.1}$$

$$\sum_{i=1}^{4} x_i = 1,$$

$$x_i = 0 or 1$$

D_{rec} is the expected recurring travel time, also known as historical travel time, TT_h that is location and time specific, based on the fact that the average speed on roads are usually similar at the same time of different days of the week. In transportation there is a big difference between weekday and weekend traffic. We are only looking at the weekday traffic which is usually worse than the weekend. Expected values are derived offline using past historical data for each segment. The preloaded digital maps available on the vehicular nodes may provide this traffic statistic of the roads at different time of the day. On the other hand, the observed travel time along a segment can be easily obtained by each vehicle in real time by its sensors. If the observed travel time on the segment is higher than a threshold, which is determined as in [10] by multiplying the congestion factor c with the expected recurring delay, the travel time is said to be excessive.

$$\forall r \in E, oTT > (1 + c) * r.TT_h \Rightarrow oTT \text{ is excessive.} \tag{3.2}$$

3.1.1.1 Monitoring of traffic variables

A fundamental aspect of our first objective is to identify the relevant microscopic and macroscopic traffic variables that connected vehicles should monitor along their trajectory in order to estimate the cause of the congestion they detect. Common causes of congestion are a widely investigated topic within the research community. Available research focus on studying the influence of a cause on traffic, such as the influence of weather, special events, incidents, workzones and bottlenecks on traffic. In order to identify a set of variables representing spatial and temporal features of the components of congestion capable of distinguishing non-recurring congestion in an efficient way, we highlight from previous research key spatio-temporal variables characterizing each cause of congestion.

Inclement weather (rain, fog, snow, ice) has an impact on the fundamental macroscopic traffic flow variables (flow, speed and volume). It was also shown that microscopic traffic variables such as desired speed, desired acceleration and deceleration and minimum following distance parameters can be influenced during snowy road conditions for different reasons [6]. The free flow speed is defined by the speed driven when the driver is not influenced by nearby road users. Free speed is reduced to desired speed if the speed is influenced by other drivers, the road, characteristics of the vehicle, conditions such as weather and traffic rules (speed limits). For instance, drivers reduce their speed in order to avoid skidding in inclement weather. A reduction of desired speed, which reaches up to 30% for snowy roads, has been found. The reaction to adverse weather conditions varies between regions. As for the

desired acceleration and deceleration, a slippery road reduces friction between road surface and tyres, thus drivers are not able to accelerate and decelerate as strongly as compared with dry road conditions, that is, maximum acceleration and deceleration decrease during snowy road conditions. Moreover, drivers reduce acceleration and deceleration to avoid skidding. Finally, drivers try to maintain a higher minimum following distance in order to cope with longer stopping distances caused by slippery roads [7]. From these findings, we assume that the features collected along a vehicle's trajectory and experience on other road segment that can help infer weather conditions are : higher minimum following distance, reduced desired speed and higher travel times on some segments of the trajectory.

Regarding special events (sport games, concerts, religious activities, political demonstrations), previous researches have highlighted that they may lead people to travel towards the same destination in a very limited time interval, and then to leave the venue again in a very short time span [8]. The impact of a special event on traffic has been thoroughly studied in [9] for different demand categories of people going to the event (inbound traffic) and leaving after its end (outbound traffic). Special events may cause congestion depending on the intensity of ingress traffic demand or sharp traffic surge in concentrated time span. Thus a special event has an impact on the traffic behaviour in a specific region over time. Such an *impact region* can be defined as the list of congested segments of the road network around the special event. From these conclusions, we also identify features along a vehicle's path that are the most informative of a special event traffic condition. Firstly, we assume that if a vehicle experiences a NRC caused by a special event, then the vehicle is necessarily in the impact region of the event. Secondly, we hypothesize that along the path of the vehicle, if some road segments of the vehicle's path are inside the impact region, and the travel time on those segments are abnormally high, then we may associate this characteristic to a congestion caused by a special event. Finally, we assess the observed demand along a vehicle's path in order to detect the presence of a sharp traffic surge. These features were found to be especially relevant to measuring operational performance during special events.

Incidents (emergencies, accidents, vehicular breakdowns, road defects, police checks) and workzones also have an impact on the traffic flow variables. Particularly, these events physically block one or multiple lanes of the road segment. The blocked section of the road is called problematic spot (Pspot) and can be defined by an interval of start and end point [s, e] depending on the lane. No vehicles can traverse this section, thus reducing the capacity of the road. Trajectories of vehicles on a segment comprising a problematic spot cannot register coordinates of the blocked section. Regarding workzones, it was shown that if they happen at day time, especially at rush hours, their impact on traffic is severe, sometimes exceptionally larger than that of traffic accidents happening at the same time. On the other hand, if they

happen at night, their impact is not that significant. Often times, workzones occupy the road segment a longer period of time than incidents because incidents are undesired and should be cleared as fast as possible. Instead of features extracted from experience on other road segments, we identify features that are specific to the road segment where the excessive delay is observed to infer an incident or a workzone event. Essentially, we suppose that if there is a Pspot on the road, then the NRC is caused by either an incident or a workzone. We also assume that the duration of the Pspot is a good indicator of a workzone event.

3.1.1.2 Features extraction

After obtaining quantitative fine-grain traffic variables, self-organized vehicular ad hoc networks can further monitor the variation in traffic information of neighbour nodes in order to estimate real-time features. We present below relevant features that can be extracted from observable trajectory data : CurrentTravelTime, TrajectorySpeed, TrajectoryGap, TrajectoryTravelTime, Pspot, ImpactRegion, ImpactRegionTravelTime, TrajectoryDemand, StoredEvent.

1. CurrentTravelTime : According to Equation 3.2, this feature categorizes the travel time observed along a segment as normal or excessive.
2. TrajectorySpeed : It is not trivial to assess the macroscopic impact of an event on the road network from a vehicle point of view. Investigation of features from a microscopic view is required at a vehicle level to better assess the situation and detect the weather condition. The TrajectorySpeed aims at summarizing speed data along the vehicles' path in order to detect the presence or absence of a weather condition. The large number of data collected is aggregated into a valuable indicator of normal or abnormal weather conditions. The deterministic aspects of vehicles movements include, following the speed of the front vehicles and driving within the speed limits. While random aspects include : changing lane, passing other vehicles, radical changing the speed because of an accident or the appearance of heavy traffic and excessive speed. According to the car-following model developed by Stefan Krauss [72], each vehicle speed can be computed as per Equation (3.3) below. The model uses the following parameters :

   $\boldsymbol{a}$ **:** the maximum acceleration of the vehicle (in m/s^2)

   $\boldsymbol{b}$ **:** the maximum deceleration of the vehicle (in m/s^2)

   v_{max} **:** the maximum velocity of the vehicle (in m/s)

***e* :** the driver's imperfection in holding the desired speed (between 0 and 1)

The safe velocity is computed using the following equation :

$$v_{safe}(t) = v_l(t) + \frac{g(t) - v_l(t)\tau}{\frac{\overline{v}}{b\overline{v}} + \tau}$$

Where :

v_l***(t)* :** speed of the leading vehicle in time t

***g(t)* :** gap to the leading vehicle in time t

τ **:** the driver's reaction time (usually $1s$)

As v_{safe} may be larger than the maximum speed allowed on the road, the minimum of these values is computed as next. The resulting speed is called the desired speed.

$$v_d(t) = \min\{v_{safe}(t), v(t-1) + a, v_{max}\} \tag{3.3}$$

It should be assumed that most drivers will operate their vehicles in a reasonable manner, considering such things as road and weather conditions, traffic volumes, adjacent obstructions and distractions. This assumption leads to the theory of 85^{th} percentile speed, defined as the speed 85% of drivers are moving at or below [73]. TrajectorySpeed is referred to as normal or abnormal, depending on the 85^{th} percentile speed the vehicle calculated along its trajectory.

3. TrajectoryGap : Similarly, this feature also aims at summarizing minimum following distances along a vehicles' trajectory. TrajectoryGap is categorized as high or normal.
4. TrajectoryTravelTime : The vast majority of drivers want to reach their destinations as quickly as possible. This feature aims at aggregating values of travel time. Since to every road segment is associated an expected travel time, we compare the observed travel time on a segment with the expected travel time. We evaluate the travel time on all segments of the path and compare them to their respective expected travel time. TrajectoryTravelTime is then the indicator of normal or abnormal traffic delay. In order to reduce the randomness of a vehicle driving voluntarily slow, the average method is taken to calculate the travel time and considers the impact of the travel time of all vehicles within the scope, the formula is set as follows :

$$TT = (1 - \alpha) * TT_s + \alpha * \overline{TT_t} \tag{3.4}$$

where TT_s is selected as the travel time of the representative, TT_t is a mean value of

the vehicular travel time in the wireless coverage of the representative vehicle and α is a weighting factor which means the different degrees of importance.

5. Pspot : This feature is specific to the road segment where the excessive delay is observed and uses position data to detect the presence or absence of a problematic spot. As in [38], we use position data to extract the distribution of vehicle footprints on the road. When the cause of the NRC is an incident or workzone, lots of vehicles periodically register coordinates of their neighbors. If a section of the road segment is blocked, no position coordinates are recorded between the start and end point [s, e] of the problematic spot as shown in Fig. 3.1. This feature also considers the temporal aspect of the observed problematic spot. It is a good indicator of an NRC caused by a workzone if the event lasts more than one hour.

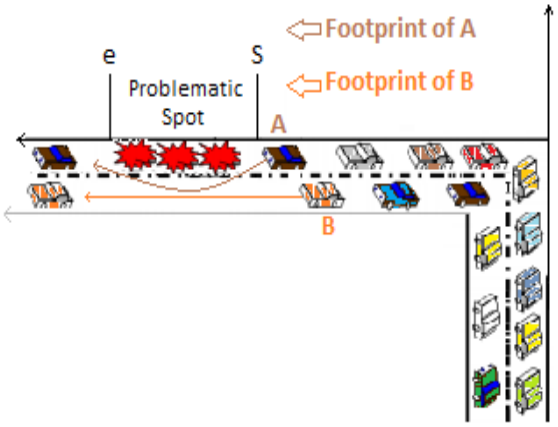

Figure 3.1 Trajectory of vehicles around a problematic spot

6. ImpactRegion : Each segment of the road network is labeled as inside or outside of an impact region. Transportation impact studies are carried out for individual developments, and traffic engineers are able to identify a list of congested segments of the road network around developments where special events occur.
7. ImpactRegionTravelTime : This feature is extracted in the same manner as the TrajectoryTravelTime except that it looks if in the path of the vehicle, there are road segments that are labeled inside an impact region. The travel time on those segments is either normal or abnormal. ImpactRegionTravelTime might also indicate that no segment of the path of the vehicle is inside an impact region.
8. TrajectoryDemand : From the average densities and speeds, we aggregate the values to obtain an understandable TrajectoryDemand feature. The observed traffic demand

along a vehicle's path is either high or normal.

9. StoredEvent : Excessive travel time can be noted on a road segment adjacent to one where the congestion was initially detected. Cooperation between vehicles can propagate the event to a maximum number of adjacent roads. StoredEvent might indicate that incident, workzone, special event, weather, bottleneck or that no stored event on the segment exists.

Finally, along each road segment, besides the features extracted above, each vehicle computes a local traffic evaluation. They estimate average speed, average density and travel time on the segment. Based on those characteristics, the congestion status can be locally identified. A vehicle can recognize via its neighbours if it's in a jam via cooperative VANET congestion detection. Numerous research studies have investigated VANET-based congestion detection [1], [19] and any of these solutions can be used to accurately detect congestion. Congestion is a particular state of mobility and as in [19], we will use mean speed, estimated traveling time and density characteristics to detect highly congested road segments along a given direction in urban areas. Following local traffic evaluation, if the segment is congested and from Equation 3.2, oTT is excessive, the observable trajectory characteristics and the results of the local traffic evaluation are provided to the feature vector as input to the classification model for inference on the cause of the NRC.

It's only when a vehicle detects congestion and that it is excessive, that the vehicle enters the inference process. The focus of the first phase is to integrate these observed traffic flow variables characterizing NRC events into the classification models.

3.1.2 Classification models

As part of the iterative process of building and refining models, we experimented with many machine learning methods in order to solve the classification problem at hand. All models require a feature vector in order to predict the output class. The feature vector x is composed of the following input variables used by the classification models for prediction :

x_1 **:** CurrentTravelTime

x_2 **:** TrajectorySpeed

x_3 **:** TrajectoryGap

x_4 **:** TrajectoryTravelTime

x_5 **:** Pspot

x_6 **:** ImpactRegion

x_7 : TrajectoryInsideImpactRegionTT

x_8 : TrajectoryDemand

x_9 : StoredEvent

All features have finite discrete domains and the dependent target variable Y that we are trying to predict is the NRC cause. The target variable can take a finite set of values :

y_1 : Recurrent

y_2 : Incident

y_3 : Workzone

y_4 : Special event

y_5 : Adverse weather

Tree models where the target variable can take a finite set of values are called classification trees (CT) [74]. Each element of the domain of the classification is called a class. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. The goal is to create a model that predicts the value of a target variable based on several input features. Each interior node corresponds to one of the input features ; there are edges to children for each of the possible values of that input feature. Each leaf represents a value of the target variable given the values of the input features represented by the path from the root to the leaf. Training data set comes in records of the form :

$$(x, Y) = (x_1, x_2, x_3, ..., Y)$$

C4.5 is an algorithm used to build classification trees from a set of training data using the concept of information entropy [74]. The purpose is to split at each node with the feature having the highest normalized information gain. The algorithm then recurs on the smaller subsets of the split. The recursion is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the prediction. We employ such an algorithm in the training of the CT described in this paper in Chapter 4. We present in the paper the details of the classification tree obtained. Also, in the paper, we show a possible extension of the CT method, random forests which is an ensemble learning method also used for classification.

We also developed a probabilistic model based on a Naive Bayesian classifier which gives useful predictions about the congestion. We sought to abstract the problem of traffic congestion classification to a consideration of probabilistic dependencies among a set of random variables, representing properties of key components features and contextual observations. Such representation helps to reduce the parameter space of the learning and inference effort

as well as provide a model of congestion classification. We set out to learn a statistical model that could provide inferences about the non-recurrent component that is causing the congestion. The aim of the Naive Bayesian classifier, as with other classifiers, is to assign a target variable to one of a discrete set of categories based on its observable features. We present in the paper in Chapter 4 the Bayesian network we propose for our classification problem. Also, in order to solve the classification problem, we experimented with a boosting technique. Boosting is a fairly recent technique in supervised learning. In the article in Chapter 4, we presented AdaBoostM1, a standard boosting scheme where diversity is created by focusing on where the existing model makes errors.

3.1.3 Real-time continuous evaluation of traffic

Since the purpose of this phase is to accurately detect and classify a NRC, we use communication between vehicles to propagate the local caption of a vehicle to the vicinity in order assess if the temporary induced traffic change related to an event can be mitigated in a short period or does the event represent a permanent change representing an NRC. To do so, in addition to BEACON messages that are regularly sent to nearby vehicles, two types of messages are implemented in our system : Extraordinary Event Request (RQ) and Extraordinary Event Response (RP). This process requires a 100% penetration rate of VANET enabled vehicles. If an NRC is locally detected, the back-propagation algorithm activates in real-time a process that shares the individual estimation made by the vehicle. The back-propagation algorithm is presented in the article in Chapter 4. In fact, according to Equation 3.2, most vehicles in the vicinity will have assessed the same deviation from normal recurrent road condition. The extraordinary event request is transmitted upstream via broadcast to all nearby cars in its communication range, and hop by hop until the entry of the road segment. RQ contains the time of the event, segment ID and event type fields. RQ escalades backwards because of congestion spill-back and the consequence of an extraordinary event will have measurable impacts on the K adjacent roads. Also, according to [10], the minimum duration before an extraordinary event is recognized as NRC corresponds to the number of consecutive link journey time that the episode contains. To this purpose, every vehicle on the segment maintains the extraordinary event request. The last vehicle on the road segment is responsible of sending it back upstream once it is going to exit the road segment in order to ensure the event lasted enough time and that it corresponds indeed to a high confidence episode. Each vehicle has 3 modes :

Mode 1 : One activates when it is ready to send an extraordinary RQ or RP event and no one has sent one yet on the segment.

Mode 2 : Another executes when the vehicle is near the virtual line of the road or at the entry of the road segment.

Mode 3 : The other is idle, listening to events : the decision of forwarding is based on the overheard retransmissions. Each node keeps track of the messages received during the time slot of t seconds, where t corresponds to the inverse of the distance to the node in the last hop.

Transmitting extraordinary events helps to obtain a more scalable traffic evaluation and detect reasonable NRC. A vehicle sends events upstream only if a reasonable NRC is detected. If a vehicle enters a road segment and receives the RQ while the duration is not yet reached, it is in mode 2. If the event is an accident or workzone, vehicles do not wait for the duration to end because the NRC is considered reasonable. An extraordinary event response is sent to first order adjacent roads. Opposite direction to the flow of traffic on adjacent segments are not connected to the segment where the event occurs. If congestion is due to weather, special event, the last vehicle has to store, carry and forward backwards the extraordinary request once it reaches the end of the road segment, to reach the duration before high confidence episode can be recognized. Similar to the RQ, the RP contains the time of the event, segment ID and event type fields where the message is defined as a response event.

3.1.4 Performance analysis

Firstly, we present in this section a detailed analysis of how each feature performed in terms of accurately being able to split the macroscopic or microscopic data collected.

CurrentTravelTime on a segment measures the travel time of a vehicle on each segment and compares it with TT_h of each segment. The observed travel time along a segment can be easily obtained by the vehicles of the VANET. If it is higher than a threshold, which is determined by multiplying the congestion factor c with the expected recurring delay, the travel time is said to be excessive, or else, it is considered normal. Fig. 3.2 shows the expected recurring travel time TT_h that is location and time specific. Values are derived from the base scenario for each segment and tabulated in a database.

Particularly, we examined the travel time variability on the base scenario to get the appropriate value of c in the urban network characterized by signalized and interrupted flow. Fig. 3.3 shows the variability of the collected values of travel times for vehicles on a particular segment. We studied different road segments with an average of 800 vehicle data point per segment and we varied the congestion factor to find that the appropriate value of c for the urban network in order not to get excessive travel time in absence of NRC in the base scenario is 1.8. This is shown in Fig. 3.3, when $c = 1.8$, most data points are below the curve.

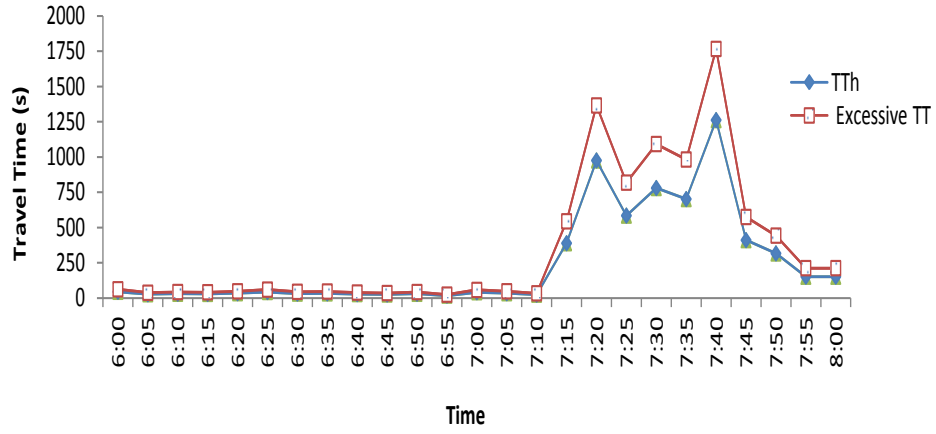

Figure 3.2 Historical TT on a segment at 5 minutes interval and excessive TT with c=1.4

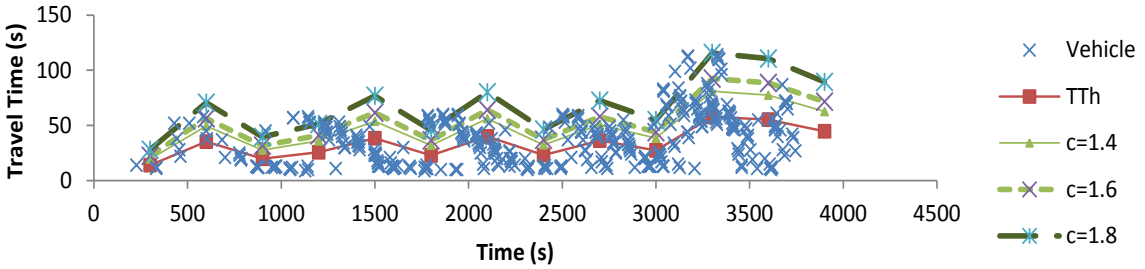

Figure 3.3 Variability of travel time data

TrajectorySpeed feature analysis is conducted starting with creation of a cumulative speed distribution curve of the desired speeds collected along a vehicle's route. Fig. 3.4 illustrates *v(t)* the speed of a vehicle on an urban interrupted road network. On the figure, we highlight the desired speeds that the vehicle is attaining along its trajectory. A frequency distribution table lists the number of vehicles observed at each desired speed. Fig. 3.5 shows the distribution in percent. This allows for a more detailed speed analysis. The 85^{th} percentile speed is shown where the plotted curve intersects the 85% line and is 14.2614 m/s for this vehicle in the base scenario. In Fig. 3.6, we show the 85^{th} percentile speed measures collected by vehicles in different scenarios. We note the effect of weather on mobility. Percentile values indicate that most vehicles lower their speeds in inclement weather.

Similarly, Fig. 3.7 illustates *g(t)*, the empty space after the leading vehicle along a vehicle's

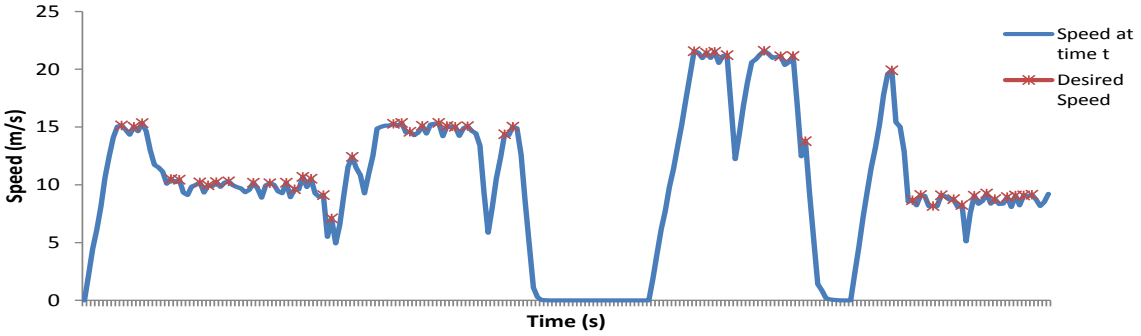

Figure 3.4 Desired speeds of a vehicle along its trajectory

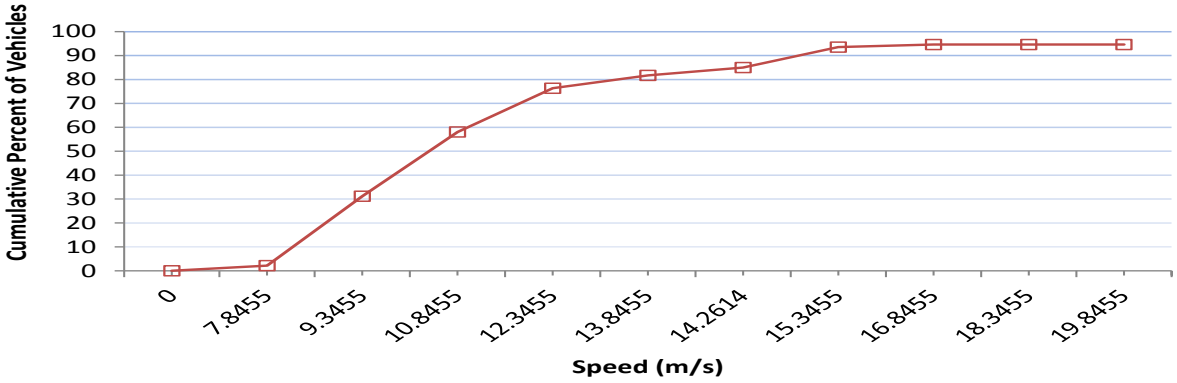

Figure 3.5 Cumulative speed distribution curve of a vehicle in the base scenario

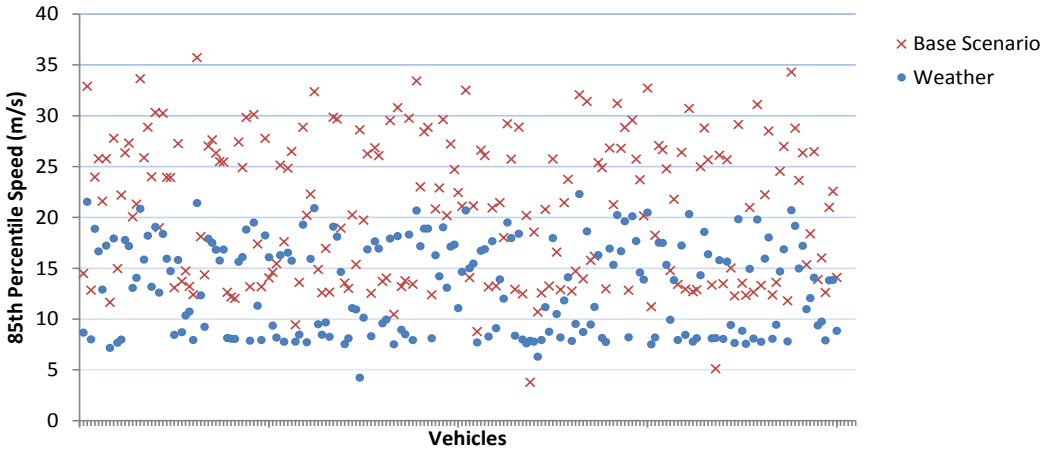

Figure 3.6 85th percentile speed of vehicles in different scenarios

trajectory in the urban network. The following distance mainly depends on the speed of the following vehicle which is adapted to the speed of the leading vehicle. The desired following distance between two consecutive vehicles is highlighted in Fig. 3.7 and corresponds to the minimum safe gap attained by the vehicle. TrajectoryGap feature analysis starts with the collection of desired gaps, creation of cumulative gap distribution and calculation of the 85^{th} percentile gap observed. In Fig. 3.8, we show the 85^{th} percentile gap measures collected by vehicles in the base scenario and the weather scenario. Incident, workzone, special event and bottleneck scenarios showed values that are similar to the base scenarios, thus, they are not presented. From the figure, we see that in inclement weather, most vehicles augment their following distance in comparison to normal weather conditions.

TrajectoryTravelTime of a vehicle measures the observed travel time of a vehicle on each segment of its route and compares it with TT_h of each segment. Fig. 3.9 shows TT_h, oTT and the excessive travel time for each segment of a vehicle's route on the base scenario. We note that in normal conditions, oTT is always lower than excessive travel time. In Fig. 3.10, we compare trajectory travel time of the same vehicle in different scenarios. From these results, we can classify the travel time along a route as normal or abnormal. Classifying the base scenario as normal behaviour, the feature shows no deviation from normal in the incident, workzone and bottleneck scenarios. Atypical behaviour is evident in the weather and special event scenarios and that is why we only present these scenarios. As indicator, we use the average percent deviation from excessive travel time.

$$\text{Average travel time deviation}(\%) = \frac{eTT - oTT}{eTT}$$

We only consider observed travel time on edges that are below the excessive travel time because the other edges will have another treatment as they are indicators of another feature. The vehicle in Fig. 3.10 experiences 66.63% average deviation of travel time in the base scenario, 43.85% in the weather scenario and 62.81% in the special event scenario. It shows that along a vehicle's trajectory in the weather and special event scenarios, oTT's are closer to excessive travel times than in other scenarios.

TrajectoryInsideImpactRegionTT considers from the TrajectoryTravelTime feature, the edges from the route of a vehicle that are inside an impact region. If travel time along those edges is near the excessive threshold, it is considered abnormal. We use as indicator a weighted

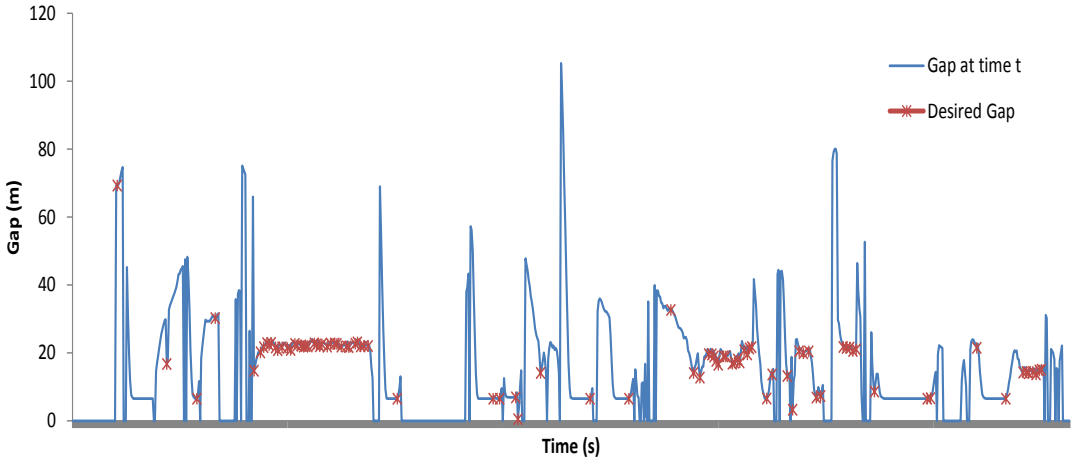

Figure 3.7 Following distance of a moving vehicle in the base scenario

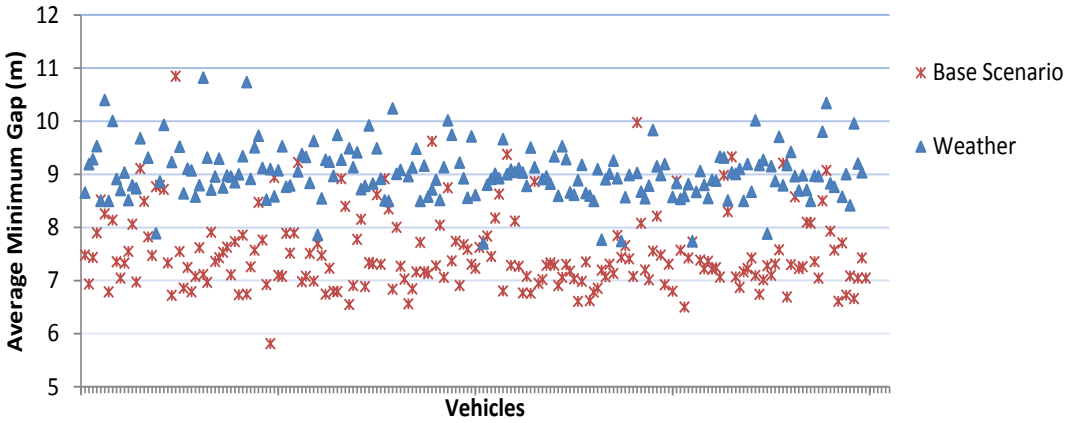

Figure 3.8 85th percentile gap values of vehicles in different scenarios

average travel time deviation (WATTD).

$$\text{WATTD}(\%) = \frac{\beta_1 \frac{oTT_1}{eTT_1} + ... + \beta_n \frac{oTT_n}{eTT_n}}{\sum_{i=1}^{n} \beta_i}$$

$$\text{with } \beta_i = \frac{oTT_i}{eTT_i},$$

and n = number of edges inside impact region.

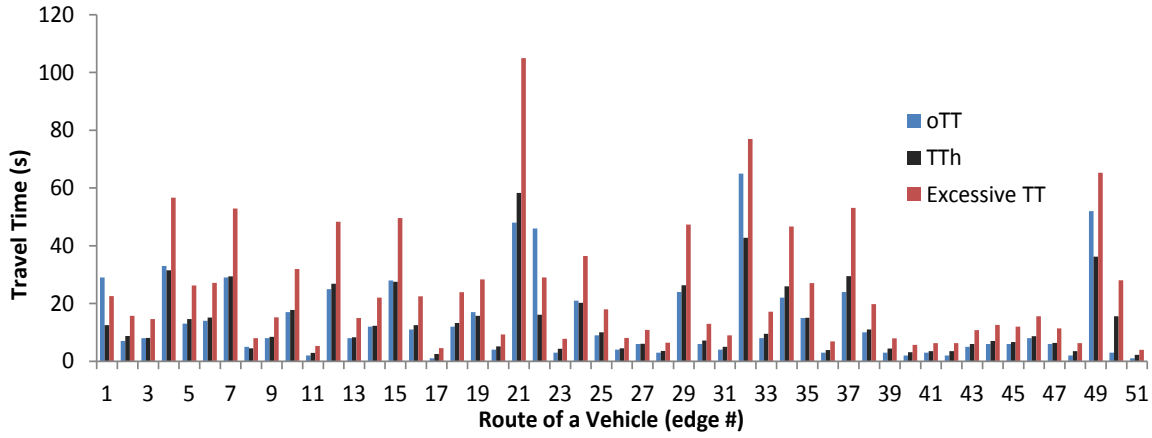

Figure 3.9 Trajectory travel time on edges of a route of a vehicle in the base scenario, c= 1.8

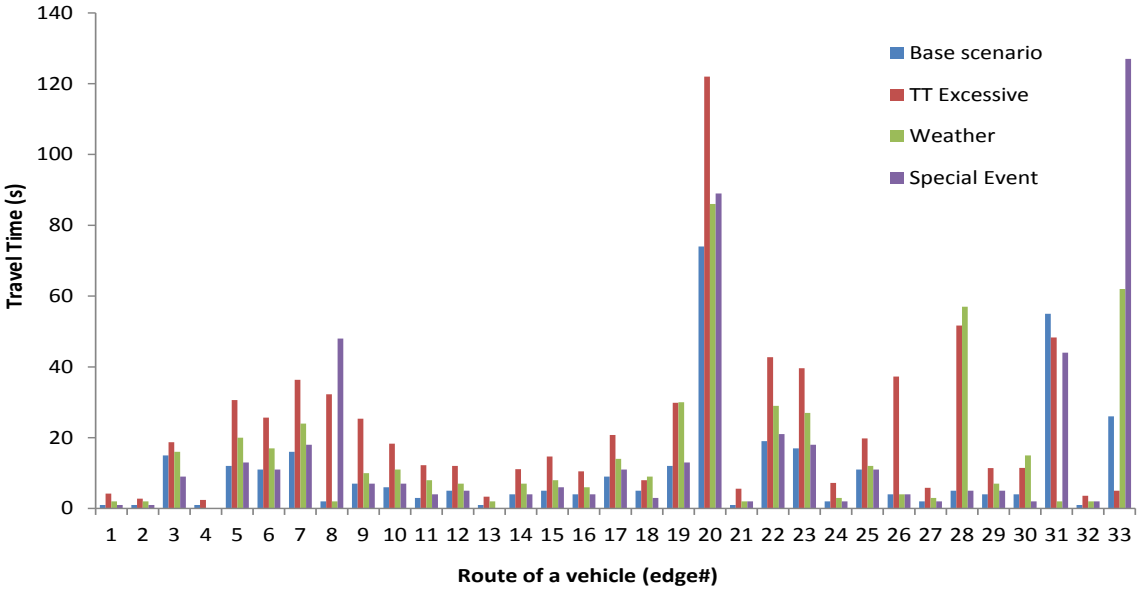

Figure 3.10 Comparative observed travel time along a route of a vehicle in different scenarios

TrajectoryDemand measures the flow on the edges of a route. We analyse below how this feature is extracted. Macroscopic stream models represent how the behavior of one parameter of traffic flow changes with respect to another. These relations are shown in Fig. 3.11, Fig. 3.12 and Fig. 3.13 for an edge in the urban base scenario. The simplest assumption is the assumed linear equation between speed and density proposed by Greenshield [75].

$$v = v_f - (v_f/k_j) * k$$

where v is the mean speed at density k, v_f is the free speed and k_j is the jam density. The relation with flow and density can be derived because the flow $q = k * v$. We get the following parabolic equation :

$$q = v_f * k - (v_f/k_j) * k^2$$

Finally, we derive the relation between speed and flow :

$$q = k_j * (v - v^2/v_f) \tag{3.5}$$

Once the relationship between the fundamental variables of traffic flow is established, the boundary conditions can be derived. The boundary conditions that are of interest are jam density, free flow speed, and maximum flow. From Equation (3.5), we find the critical density at maximum flow by the following derivative :

$$\frac{dq}{dk} = v_f * (1 - 2 * (k_c/k_j)) = 0$$

$$k_c = k_j/2$$

Therefore, density corresponding to maximum flow can be approximated by half the jam density. Once we get k_c, we can derive maximum flow, q_{max}.

$$q_{max} = (v_f * k_j)/4 \tag{3.6}$$

Thus the maximum flow is approximated by one fourth the product of free flow and jam density. Finally, speed at maximum flow, v_c, is found by substitution.

$$v_c = v_f/2$$

Therefore, speed at maximum flow is half of the free speed. Each edge in the simulation of urban mobility with SUMO has one of three speed limit values assigned to it, $v_f = [8.33, 13.89, 19.44]m/s$. Also, the space occupied by a vehicle is approximately $6.65m$. Jam density on a segment corresponds to the maximum number of vehicles on a segment divided by the number of lanes times the length of the segment. Considering this approximation, $k_j \cong 0.155038V/m$. Using the speed-flow relationship and knowing the average observed speed on the segment and current density, we estimate flow and compare flow with q_{max}. Demand is normal or high. This detailed breakdown of demand based on average speeds and density is useful to reflect local demand that significantly characterize traffic condition on an edge. Fig. 3.14 shows the demand in terms of flow along a vehicle's trajectory in different

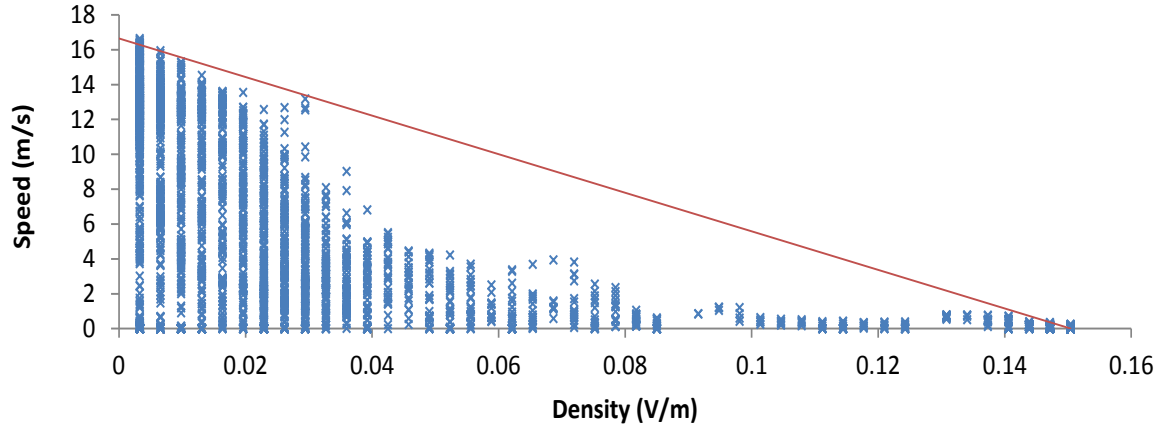

Figure 3.11 Speed-Density

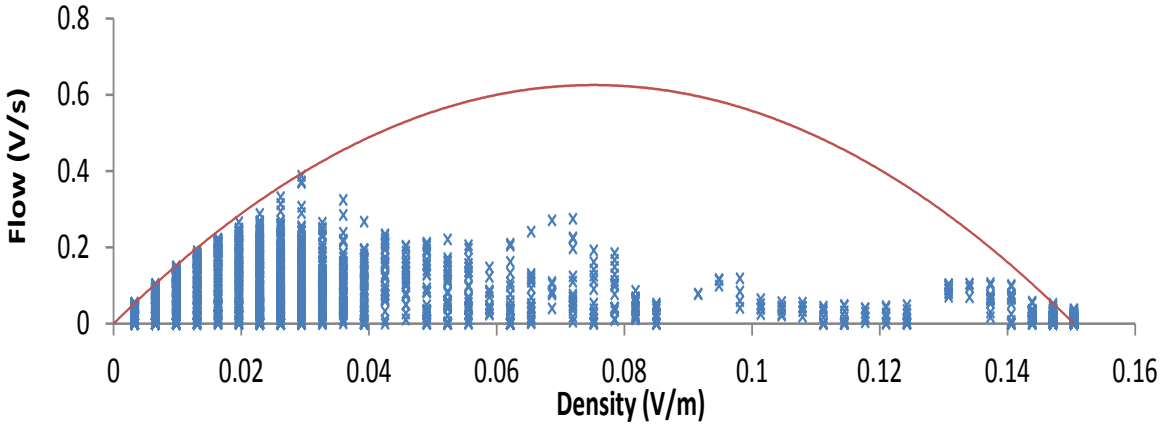

Figure 3.12 Flow-Density

scenarios. We also use the weighted average as indictor of excess demand along a trajectory. Results indicate that weather scenarios show lower weighted average demand than other scenarios. TrajectoryDemand is particularly high in the special event scenario. Special event related traffic has a much sharper traffic surge ; exceeding threshold in a shorter, concentrated time span. Traffic demand distribution in all other scenarios appears normal compared to the q_{max} denominator. Since the impact depends on the existing condition of the particular road that is being impacted by the change in demand, we only use this feature as overall sense and feel of the road segment in comparison to a denominator q_{max} that makes this measure comparable.

We show in Fig. 3.15 and Fig. 3.16 how the parameters of traffic flow change with time in

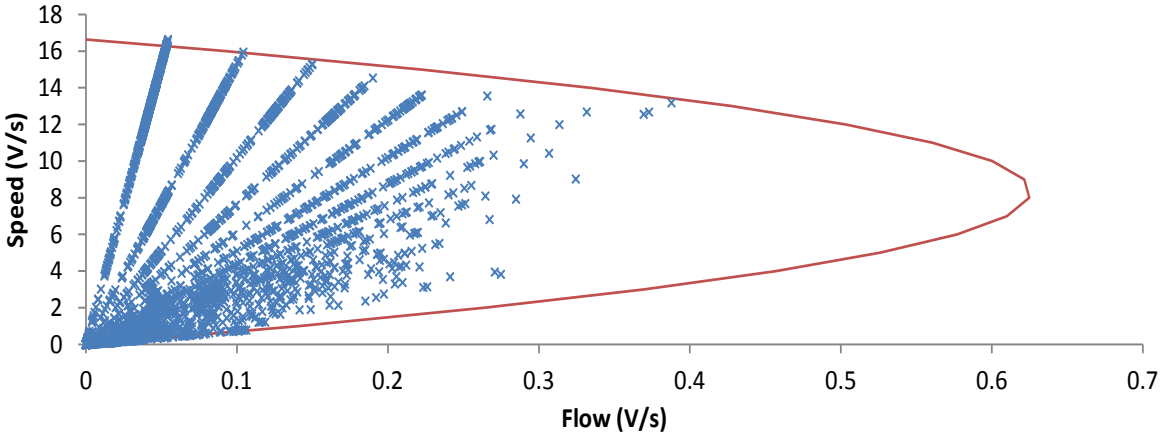

Figure 3.13 Speed-Flow

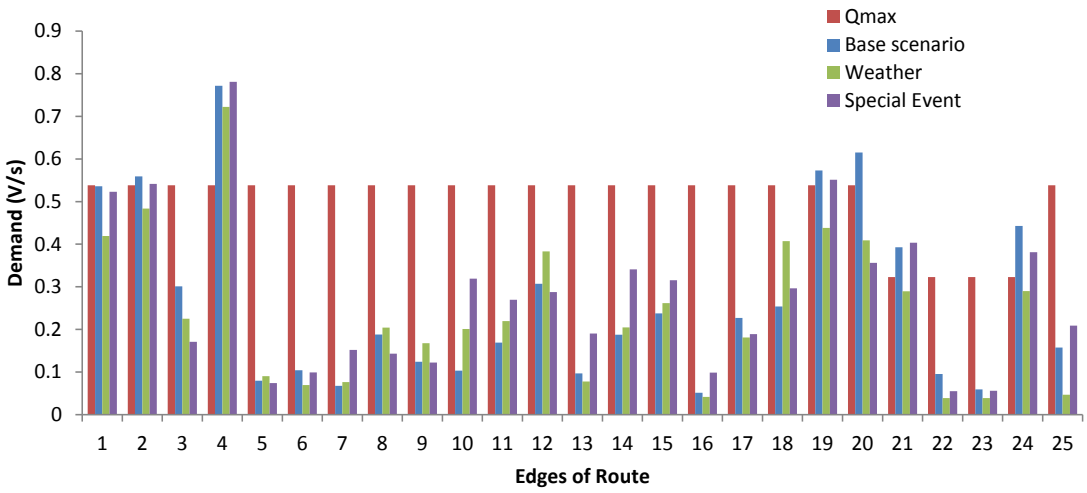

Figure 3.14 Trajectory demand along a route for different scenarios

the urban road network on a particular edge. The flow is zero either because there are no vehicles or there are too many vehicles so that they cannot move. We observe mean speed and density on a segment in order to declare congestion status. It's only when vehicles detect congestion that they analyze if it is excessive. In case oTT is excessive, vehicles assume they are experiencing NRC and collect the observable trajectory characteristics in order to form a feature vector for the classification models.

Finally, to evaluate the performance of the classification models, the accuracy of the classifier is the primary metric and is determined by the percentage of the test dataset examples that are correctly classified. For the CT, we performed 10 fold cross-validation on the training set

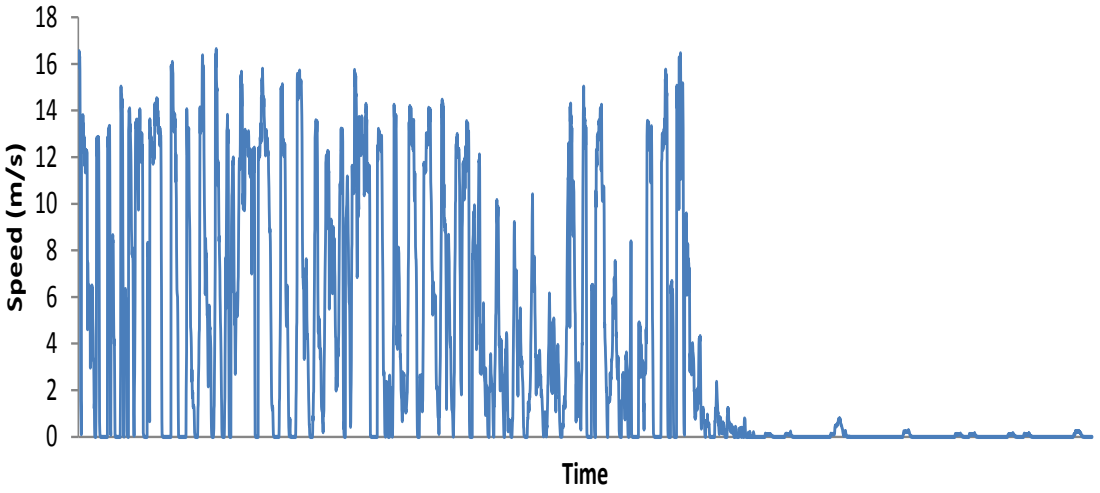

Figure 3.15 Variation of mean speed on an edge

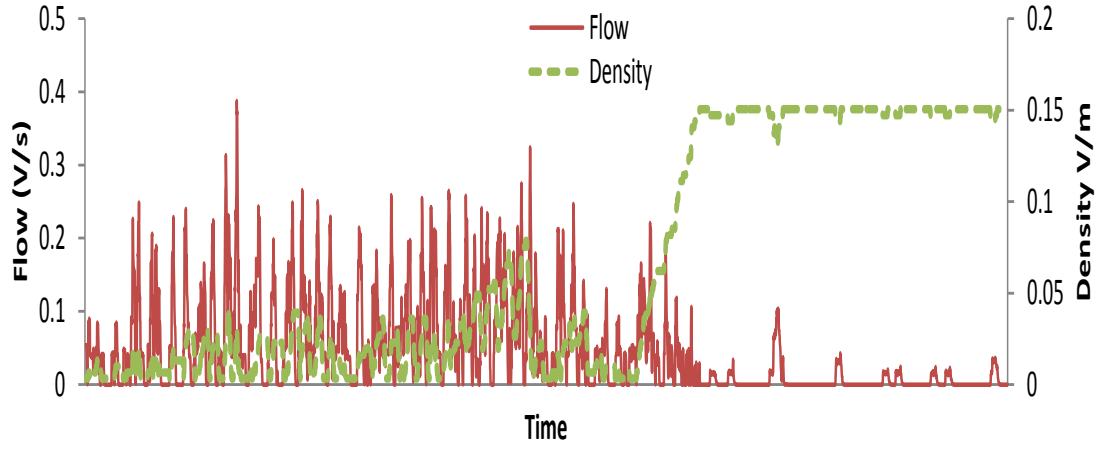

Figure 3.16 Variation of flow and density on an edge

and we got 87.98% of correctly classified instances. For the Naive Bayesian classifier's, we got 88.83% of correctly classified instances. Weka's implementation yielded 89.51% accuracy of Random forest of 100 trees, each constructed while considering 4 random features. And, AdaBoostM1 with 10 iterations, yielded 89.17% of accuracy.

3.2 Phase 2 : Cooperative evaluation of the cause of congestion

Because traffic is multifaceted and to conceal the fact that individually, vehicles have partial knowledge about the road condition, in the second phase of our work, we implemented an evaluation process to increase even more the estimation accuracy of the classifiers regarding the cause of congestion. Unlike the back-propagation algorithm presented in the previous section, vehicles not only propagate information, we show that if they cooperate, they evaluate better the traffic situation and thus increase estimation accuracy. To do so, we needed to simulate scenarios extended from a realistic urban city vehicular motion traces in order to build a synthetic dataset to feed the mining techniques we propose for learning purposes. The completion of objective 3 was carried out in this phase. The objective was attained as described in the article titled "Cooperative Evaluation of the Cause of Urban Traffic Congestion via Connected Vehicles" presented in chapter 5.

3.2.1 Data mining methods

Firstly, congestion in an urban network is mainly caused by incidents, work-zones, special events, adverse weather, or recurrent congestion. On the other hand, vehicles are equipped with a method to detect excessive congestion in an urban network and from the work in the previous phase, they have a classification algorithm on board able to attribute a possible cause to it. In this phase of the study, we seek to improve the vehicle's estimation of the cause of congestion and reduce false alarms by cooperative methods infused with knowledge about the others evaluation specifically in the event of traffic congestion. In fact, the classification algorithm returns the cause with the greatest probability, the most likely cause. We make use of the other probabilities computed by the classifier to extract more knowledge. We propose that each vehicle represent its uncertainty about the cause of congestion in a vector of probabilities associated to each of the possible causes of congestion. In particular, the vector of probabilities exchanged between the CVs should have this form :

C = [P$incident$, P$workzone$, P$weather$, P$specialevent$, P$recurrent$]

After proper representation of vector C, different methods to elaborate a decision concerning the cause of urban congestion on the segment are presented in the paper in Chapter 5. The result is a voting procedure, belief functions method and a data association technique. Each vehicle has its own decision module and it contains one of the Voting Procedure (VP), Belief Functions (BF) or Data Association Technique (DAT).

3.2.1.1 Voting procedure

When a vehicle experiences excessive congestion and wants to evaluate the cause of the urban traffic congestion on the road segment, the probabilistic classification model on board of each vehicle predicts the cause of congestion and returns the result in the form of the probabilities vector presented above, with one cause of congestion having the highest probability. The vehicle broadcasts the probabilities vector in the vicinity. Vehicles on the road segment collect the messages received and the decision module on board of each vehicle computes the counts for each cause. The cause having the highest count is highlighted by this voting procedure as being the cause of congestion on the road segment.

3.2.1.2 Belief functions

The theory of belief functions [39] extends both the Set-membership approach and Probability Theory. The aim of this theory is to improve the level of knowledge and thus enhance the prediction accuracy of the cause of congestion experienced by vehicles on the road segment. Firstly, vehicles detect excessive congestion and are able to estimate the cause with the classification models introduced in the paper of Chapter 4. If the most likely cause of congestion computed by a vehicle is an incident, then the model can also inform that the second best possible cause it predicted is a weather condition. In this case, there is a singleton {Incident} and a 2-items subset, {Incident, Weather}. We transfer the Bayesian probability of the subset to a mass function. A mass function m is held by each vehicle and is defined on the frame of discernment Ω={Incident, Workzone, Weather, SpecialEvent, Recurrent}. Each vehicle assigns a mass on any of the singletons and another on a subset containing the singleton. The subset containing the singleton represents added knowledge about the sensed traffic condition. We limit the strategy to 2-items because results showed very little improvement in the accuracy of prediction when more items are considered. The strategy is detailed in the paper, and we present how the theory of belief functions is applied to our problem.

3.2.1.3 Data association technique

This method also aimed at improving the level of knowledge from exchanged messages for efficient evaluation of the cause of congestion. But this time, since each vehicle's assessment is communicated to vehicles in the vicinity, we collect the vector of probabilities exchanged by the vehicles in many scenarios to build a supervised dataset. The association rule mining method is presented in the paper of Chapter 5. Association rules are *if/then* relationships that help uncover seemingly unrelated data in a relational database. We analyse the messages

for frequent patterns in order to identify the relationships for rule generation. We extract the general association rules from the messages exchanged regarding the cause of the congestion. Consequently, data association between messages exchanged by the connected vehicles helped further scrutinise the road condition.

3.2.2 Synthetic dataset for training

In order for any machine learning model to learn anything, we need a dataset for training. However, because the technology of connected vehicles is at its early stage, there is no dataset of vehicles exchanging data in a V2V setting available for research. Our strategic contribution was the construction of a synthetic dataset for training. Synthetic datasets are designed to obtain information via simulation, while still maintaining statistical properties of the original data. Our experiments utilize validated real-world traffic traces of the Travel and Activity PAtterns Simulation (TAPAS) Cologne scenario because the realism of the simulation is a paramount aspect in transportation engineering ; traffic traces are available for research in civil engineering and they report the coordinates of each vehicle on the map every second. TAPAS Cologne scenario is assumed to be one of the largest traffic simulation data set [45] as it covers the main road network within the inner city of Cologne. The information is generated from a combination of various different sources including Floating Car Data, GSM probe data and data from stationary sensor obtained from local traffic management centers. From these combined sources, a base scenario is derived. In the base scenario, traces for the 6-8am peak hours are provided. We create extended scenarios mounted on top of the base scenario to model atypical traffic conditions such as weather, incident, workzone, special event. Our experiments are then built on the extended scenarios. Evaluation of our framework using complex real-world scenarios allow determining whether the proposed models can handle the real life's complexity.

We create extended scenarios using SUMO, a microscopic traffic simulator for the simulation of urban mobility [72]. Our investigations need the microscopic view for different reasons. Fine microscopic simulations model each vehicle explicitly and compute the traffic flow's progression by modelling each vehicle's speed and lane choice, mostly using discrete time steps of one second, calculating different traffic specific values like the amount of vehicle in a specific point and so on. Also, simulating a large area is necessary in each scenario because trajectory data along a vehicles' trip needs to be collected. Finally, SUMO enables generation of trace files that are necessary for the simulation of communication in a VANET in the network simulator ns-2 [76]. In a simulation, atypical traffic conditions are not direct model parameters but must be converted into ones. We describe below the extended scenarios

of atypical traffic conditions simulated using SUMO :

1. Extended Scenario of an incident : On the base scenario, we stop two or three vehicles, for a specific amount of time, on a lane to simulate incidents. We simulate incidents at the beginning, middle and end of a lane. We also simulate incidents on different lanes, for a long or short duration as well as inside or outside of an impact region of a special event.
2. Extended Scenario of a workzone : Similar to the above extended scenario, we stop vehicles on an edge to simulate a workzone. We vary the position on the edge and the duration.
3. Extended Scenario of bad weather : We convert the base scenario into an extended scenario of bad weather, snow for example. Snow might lead to slippery roads and reduced sight, leading to decreases in the vehicles' velocities and a more careful and defensive driver behaviour. Such behavioural changes would be reflected in simulation parameters, such as the driver's preferred velocities. Parameters of the car-following model are affected by the weather. Therefore we use eWorld because it supports these events [77]. eWorld is a framework to import mapping data, visualize it, edit and enrich it with events or annotational attributes and pass it to traffic simulators. eWorld uses constructs of SUMO to simulate events. How much an event influences the speed limit is defined within eWorld, it takes the environmental characteristics into account.
4. Extended Scenario of a special event : To generate trips to a particular destination edge where there is a special event, we have to generate random departures and random routes. We use a Poisson process to generate random timings for trips. Departures will occur individually, stochastically independent to all the others in the road network, at random moments. The rate parameter λ is the demand per second from different sources in the network, and can be seen as the flow. To generate random routes, given trips are assigned to respective fastest routes according to their departure times and a given travel-time updating interval by SUMO's traffic assignment model.

We then perform experiments on these extended scenarios based on real-world traffic. The experiments are described in the article of chapter 5. Vehicles collect macroscopic and microscopic traffic variables along their trajectory. They form the feature vector. In this phase of the study, we call the feature vector a transaction for the data mining methods. Along with the supervised target variables and characteristics, we construct a synthetic training dataset. The training dataset is a matrix with rows corresponding to transactions and columns to items. The data set generation steps are shown in Fig. 3.17.

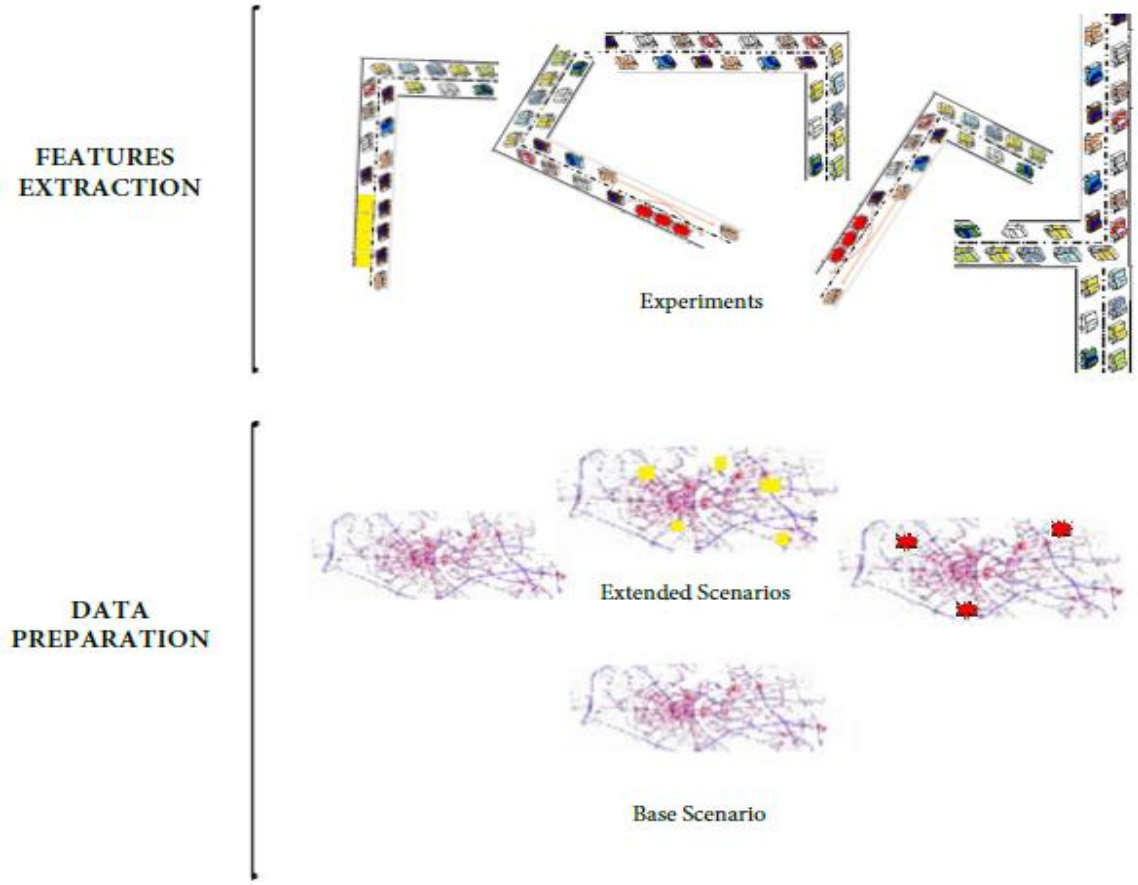

Figure 3.17 Synthetic training data set generation for model building

3.2.3 Performance evaluation

The methods are validated by three indicators ; accuracy of prediction, detection time and percentage of false alarms. When compared to the Back-Propagation (BP) algorithm proposed in the literature, the VP outperforms the BP method in terms of percentage of vehicles accurately detecting the cause of congestion. On average, it did so by approximately 48%. Also, VP outperforms BP in terms of detection time because the algorithm of the BP requires that vehicles exchange their evaluation only if they experience the excessive congestion for a certain duration of time and they are in the communication range of each other. However, we found that on average the percentage of false alarms triggered by the VP is 3% to 11% higher than the BP. A false alarm is a vehicle initiating a VP or a BP method and the simulation shows no excessive congestion, i.e. the situation captured by each vehicle is compared with the real simulation.

On the other hand, the method using Belief functions increases estimation accuracy and de-

creases detection time. Compared to the VP method, BF also decreases the percentage of false alarms by approximately 1.8%. In the recurrent scenario, BF yields the same performance as the BP method. This shows that in the evaluation process, not only cooperation between vehicles but adding knowledge to the messages exchanged improves the performance. Nonetheless, BP still outperforms BF by approximately 4.25% less false alarms in the incident, weather, work-zone and special-event scenarios. As a solution, we add more knowledge on board of each vehicle by implementing the data association technique. In the VP and BF methods, vehicles decide cooperatively without applying the association rules. When applying the mining rules in DAT, performance is greater in terms of estimation accuracy. BF models partial knowledge allowing earlier detection of the cause and DAT gives further precision on incoherencies in the data having the best estimation. In the DAT experiment, vehicles make use of the belief functions and association rules to estimate the cause of congestion. The β-DAT is detailed in the paper of Chapter 5 and improved estimation accuracy by approximately 70% compared to the BP method. Also, detection time of β-DAT is 7.09% lower than that of the BP method, informing of the congestion cause earlier. β-DAT has the lowest percentage of false alarms in all scenarios. In fact, compared to the DAT method, β-DAT decreases the percentage of false alarms by approximately 3.6%. Also, β-DAT outperforms BP by approximately 1.25% less false alarms triggered by the network on the road segment. This shows that adapting the duration in combination with cooperation between CVs and knowledge on board of each vehicle improves overall performance for the accurate estimation of the cause of congestion.

3.3 Phase 3 : Prediction of traffic flow in an urban traffic network

In the last phase of our work, our main goal was to propose a traffic flow prediction framework taking into account historical flows as well as innovative features, such as real-time reports from connected vehicles and travel time along a trajectory for accurate forecasting of flow in an urban network. In the framework, we make use of the classification model presented in phase 1 and the cooperative evaluation technique obtained in phase 2 ; this permits vehicles to classify cooperatively the cause of any congestion encountered along their trajectory and constitutes one of the innovative features to be fed to the models of prediction that we propose. In this context, the prediction of traffic flow was modeled as a multitask problem because we conjecture that when the tasks involved in multitask are semantically connected a larger improvement in accuracy of prediction can be obtained. This work is the result of the last objective of the thesis and is described in the article titled " Prediction of traffic flow via connected vehicles" presented in chapter 6.

3.3.1 Problem definition

The traffic flow prediction problem can be stated as follows. Let *Xi(t)* denote the observed traffic flow quantity during the tth time interval at the ith observation location in a transportation network. Given a sequence of observed traffic flow data, $i = 1, 2, \ldots, m$, and $t = 1, 2, \ldots, T$, the problem is to predict the traffic flow at time interval (t+Δ) for some prediction horizon Δ. This is the short-term traffic flow prediction problem. Some other works may focus on predicting the traffic flow of the next several time intervals from $T + \Delta + m$ to $T + n$ as well, it is called the long-term traffic prediction. On the other hand, most models in the literature predict flow *Xi (t+Δ)* at time *(t+Δ)* based on the traffic flow sequence $X = \{Xi, t/i \epsilon O, t = 1, 2, \ldots, T\}$ in the past, where O is the full set of observation points (roads and stations). In supervised learning, our problem becomes, given the feature X and task Y pairs obtained from history traffic flow *{(X1, Y1), (X2, Y2), . . . , (Xn, Yn)}*, learn the best parameters for a predicting model that minimizes a loss function.

3.3.2 Proposed approach

In our work, we consider the short-term traffic flow prediction problem. Since traffic flow prediction not only depends on historical data but heavily depends on real-time traffic data, we incorporate to the input feature X, not only previous traffic flows observed on the target road segment but knowledge acquired from related roads. From well engineered features, such as real-time reports from connected vehicles and travel time along a trajectory, the model learns a representation that takes into account the various events that vehicles realistically encounter on the segments along their trajectory. Moreover, the problem of predicting short-term flow is handled as a classification task. In fact, we propose that the target variable Y represent multiple classes of discrete interval of flows and the task is to predict the range of flow that the current traffic situation will generate at a near future time.

In machine learning, we normally break a complex problem down into tractable sub-problems, and learn to solve one problem at a time. This potentially ignores rich sources of information found in the training signals of other tasks. It is possible to jointly train a general system for solving different tasks simultaneously. The classifier will prefer hypotheses that explain more than one task, improving generalisation. In [78], they proposed multitask learning (MTL) as a means of inductive transfer between tasks. The update is done with both error signals. Also, a transportation system is a highly correlated network. The characteristics of transportation systems, such as the large amounts of data and the high dimensions of features, would make deep learning a promising method for transportation research.

We proposed a feedforward neural network or Multi-Layer Perceptron (MLP), which is a series of logistic regression models stacked on top of each other, with the final layer being another logistic regression because we are solving a classification problem. The purpose of the hidden units is to learn non-linear combinations of the original inputs. We can easily extend the MLP to predict multiple outputs in order to do multitask learning. Precisely, to use MTL for time series prediction, we use a single net with multiple outputs, each output corresponding to the same task at a different time. If output k referred to the prediction for the time series task at time Tk, this net makes predictions for the same task at three different times. The output used for short-term flow prediction would be the middle one so that there are tasks earlier and later that the model trained on. In particular, we propose that given a fresh new road network traffic situation at time t, Xt, the first task consists in determining what flow $c \epsilon Y$ is a suitable flow prediction at $t+5$. The second task is to find what flow $c \epsilon Y$ is a suitable short-term flow prediction at $t+15$ based on the similar road network traffic situation and on the relevant prediction of the first task and the third task is to find the flow at $t+20$.

3.3.3 Performance evaluation

The results show our approach significantly outperforms existing approaches that do not adapt to the varying traffic situations. Meaning that our models can predict flow in the presence of an incident, inclement weather, work zone or a special event. Also, since rush hours happen at almost same time of that particular day, we can even predict the flow changes at the boundary of rush hours, something the actual methods fail to accurately do. In fact, to measure the predictive power of the proposed MTL model, we compared it with the performance of the state-of-the-art ARIMA time series approach and with baseline classifiers. Baselines include : (i) RF : Random Forest ; (ii) ANN : Artificial Neural Net. MTL and the ANN model are implemented using Torch 5 package. Random Forest was from the Weka.

We use the performance index Root-Mean-Square Error (RMSE) which gives the score of the actual and predicted traffic flows at time t. We use this to measure the linear score that averages the error with the same weight and to measure the residuals by assigning larger weights to larger errors. We feed ARIMA the original traffic flow data. When compared with ARIMA and Random Forest, MTL performs best. It presents an average RMSE equal to 0.05. We compare the performance of single task learning with ANN, learning just one task of 15-min traffic flow prediction task, and multitask. Indeed, our experiments with the data show that ANN have a lower performance (0.113 for RMSE) than MTL, but higher performances than ARIMA. The results confirm the merit of MTL to forecast traffic flow and confirm the

importance of having highly related tasks.

3.4 Conclusion

This chapter presented a complete description of the research phases carried out in this thesis. The different methodologies that we adopted to solve the defined research problems were elaborated and the connection between the declared objectives and the contributions presented in the following chapters was established.

CHAPTER 4 ARTICLE 1 : DISTRIBUTED CLASSIFICATION OF URBAN CONGESTION USING VANET

Ranwa Al Mallah, Alejandro Quintero, and Bilal Farooq
IEEE Transactions on Intelligent Transportation Systems, vol. 18, issue 9.

Abstract

Vehicular Ad-hoc NETworks (VANET) can efficiently detect traffic congestion, but detection is not enough because congestion can be further classified as recurrent and non-recurrent congestion (NRC). In particular, NRC in an urban network is mainly caused by incidents, workzones, special events and adverse weather. We propose a framework for the real-time distributed classification of congestion into its components on a heterogeneous urban road network using VANET. We present models built on an understanding of the spatial and temporal causality measures and trained on synthetic data extended from a real case study of Cologne. Our performance evaluation shows a predictive accuracy of 87.63% for the deterministic Classification Tree (CT), 88.83% for the Naive Bayesian classifier (NB), 89.51% for Random Forest (RF) and 89.17% for the boosting technique. This framework can assist transportation agencies in reducing urban congestion by developing effective congestion mitigation strategies knowing the root causes of congestion.

4.1 Introduction

Congestion can be classified as recurrent and non-recurrent. Recurrent congestion refers to congestion that happens on a regular basis and usually occurs when a large number of vehicles use the limited capacity of the road simultaneously. Non-recurrent congestion (NRC) in an urban network is mainly caused by incidents, workzones, special events and adverse weather [5].

Recently, the intelligent transportation system research has shifted its focus to the next generation sensing technology, vehicular ad-hoc network (VANET). Advances in vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) wireless communications have increased the potential of real-time monitoring of traffic variables in a distributed manner. Distributed monitoring refers to the process by which macroscopic and microscopic traffic variables are collected by vehicles themselves without the need of infrastructure. A large amount of

works have been proposed for the congestion detection problem using VANET [19]. These approaches only detect congestion and do not clarify if it is due to recurrent or NRC. They cannot be used to classify the congestion into its components.

In transportation, although understanding how much of the total congestion is due to NRC has been thoroughly studied for both highway [25] and urban traffic [5], several unresolved problems still exist. Firstly, the duration, timing and location of NRC in an urban road network varies highly. Thus making it difficult to monitor traffic in real time or on a continuing basis with conventional induction loops, cameras and floating cars mechanisms which are expensive to deploy and maintain for large coverage areas. Alternative cost-effective and flexible solutions are needed to guarantee better monitoring of road traffic at various level of granularity. Secondly, existing NRC detection methods not only need extensive datasets, but they are also not deployed in real-time. In real-time, valuable information with regard to the impacts of the detected NRC can be disseminated to drivers and traffic management centers so that appropriate proactive strategies for recovering traffic conditions back to normality can be set in place. Finally, NRC detection methods should be able to better characterize a NRC event once it is detected. Existing methods only quantify the spatial and temporal impact of the detected NRC. We should be able to also classify the root cause of the NRC.

This study considers a set of unique features for each type of NRC and extracts such features from the data to infer the NRC. Thereafter, machine learning models are used to identify the specific type of NRC. Specifically, incidents and workzones are essentially characterized by problematic spots. For inclement weather, we assess the trajectory travel time, speed and gap. And special events are characterised by their impact region and demand surge. The contributions of this paper are :

— Evaluation of machine learning methods for the classification of congestion into its components taking traffic features into account for the inference.
— An algorithm able to detect, identify and propagate via VANET the cause of NRC.
— Validation of the inference methods is made relying on simulation scenarios extended from the real-world Cologne scenario [45].

This paper is organized as follows. Related work is provided in Section 4.2. In Section 4.3, we present our framework. In Section 4.4, we describe the simulation and provide results. Finally, conclusions and future work are outlined in Section 4.5.

4.2 RELATED WORK

Lots of works in congestion detection via VANET utilize machine learning to classify the traffic state into congested or free-flow [23]. To classify the level of congestion, [24] proposes a traffic congestion quantification process based on fuzzy theory. The level has values ranging from free flow to severely congested. These approaches only detect congestion and do not clarify if it is due to recurrent or NRC. The monitoring done by the schemes does not allow summarizing valuable knowledge in an efficient way.

Context-awareness is the potential to access available semantic information such as time, location, weather, temporary events and other attributes [33]. The context information used in [32] fuses different data-sources (internal sensors, web services or passenger sensors) for congestion detection. Their scheme requires additional infrastructure and communication. Without the use of external data sources for inference, our local and self-organized method classifies based on the real-time relevant information extraction by taking advantage of the streaming differentiating characteristic of VANET. Vehicles need to be context aware and able to consider multiple but adequate explanatory sources, well-tailored information won't always be available, particularly in dynamic urban networks. Due to real-time constraints much more information extraction techniques are needed to extract transport-relevant parameters. Statistical inference and machine learning algorithms can provide crucial help in this process. Understanding the causes of urban congestion is a prerequisite for deriving policies and management plans so that appropriate proactive strategies can be set in place.

4.3 GENERAL PROCESS

The observed travel time of a vehicle (oTT) along a road segment may be composed of recurrent delay (D_{rec}) and non-recurrent delay (D_{n-rec}) such as incident (D_i), workzone (D_{wo}), weather (D_{we}) or special event (D_{se}) induced delays.

$$oTT = D_{rec} + (D_i \vee D_{wo} \vee D_{we} \vee D_{se}) \tag{4.1}$$

D_{rec} is the expected recurring historic travel time TT_h that is location and time specific. The observed travel time along a segment can be easily obtained by the vehicles of the VANET. If it is higher than a threshold, which is determined as in [10] by multiplying the congestion factor c with the expected recurring delay, the travel time is said to be excessive.

$$oTT > (1 + c) * TT_h \Rightarrow oTT \text{ is excessive.} \tag{4.2}$$

We claim that real-time traffic flow data collected along a single vehicle trajectory, experience on other road segment and aggregated values offer statistically understandable spatial and temporal features that can help infer the component causing the excessive delay. The resulting classification problem takes the real-time estimates as input feature vector for inference on the cause of congestion. Thus, the general process of our framework is divided into three phases : Features extraction, classification models and cooperative process.

Phase 1 : Features extraction

A vehicle can recognize via its neighbours if it's in a jam via cooperative VANET congestion detection. The communication characteristics of a VANET are mostly based on a message called BEACON, transmitted by each vehicle every 0.1 seconds. The message contains time-stamped basic vehicle state information, such as senderID, position, direction, current speed, with optional information also possible. We present below relevant features that characterize each NRC component.

Incidents and workzones are essentially characterized by problematic spots, Pspot. As in [38], we use position data to extract the distribution of vehicle footprints (i.e., the geographical position at each sampling time point on the road). Vehicles periodically register coordinates of their neighbors. If a section of the road is blocked, no position coordinates are recorded between the start and end position of the Pspot as shown in Fig. 4.1. This feature also considers the temporal aspect of the observed problematic spot. It is a good indicator of an NRC caused by a workzone if the event lasts more than one hour. Often times, workzones occupy the road segment a longer period of time than incidents because incidents are undesired and should be cleared as fast as possible.

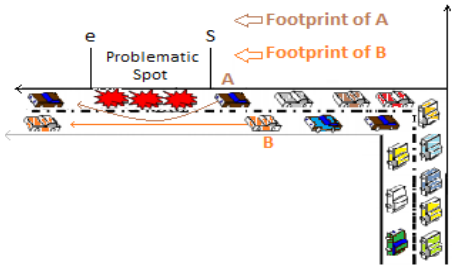

Figure 4.1 Trajectory of vehicles around a Pspot

NRC caused by inclement weather impacts the trajectory travel time (TrajectoryTravelTime), speed (TrajectorySpeed) and gap (TrajectoryGap). TrajectoryTravelTime measures travel time on the edges upstream in the trajectory of the vehicle and compares them to their

respective expected travel time. TrajectorySpeed aims at summarizing speed data along the vehicles' path. TrajectoryGap collects minimum following distances because in inlcement weather drivers try to maintain a higher minimum following distance in order to cope with longer stopping distances caused by slippery roads.

Special events are characterized by their impact region (ImpactRegion) and demand surge (TrajectoryDemand). Each segment of the road network is labeled as inside or outside of an impact region. Such an *impact region* can be defined as the list of congested segments of the road network around the special event [9]. We assume that if a vehicle experiences a NRC caused by a special event, then the vehicle is necessarily in the impact region of the event. TrajectoryDemand measures the flow on the edges of a route. Using the speed-flow relationship and knowing the average observed speed on the segment and current density, we estimate flow and compare it with maximum flow, approximated by one fourth of the product of free flow and jam density. When the vehicle detects excessive congestion on a segment, the algorithm ignores the information broadcasted from the trajectory in the last congested segments leading to the excessive congestion segment. The algorithm then only takes into account the flow on the prior segments totaling on average 1.5km. On those upstream segments, traffic is free-flowing and the flow equals the demand. We then compute the weighted average as indictor of demand along a trajectory. Since the demand surge depends on the existing condition of the particular road that is being impacted by the change in demand, we use this feature as an overall sense and feel of the road segment in comparison to the maximum flow denominator that makes this measure comparable. The procedure starts with every vehicle measuring the flow on each road segment of its trajectory.

Finally, CurrentTT is a feature that categorizes the travel time observed along a segment as normal or excessive according to Equation (4.2). Also, excessive travel time can be noted on a road segment adjacent to one where the congestion was initially detected. Cooperation between vehicles can propagate the event to adjacent roads. A StoredEvent feature might indicate that incident, workzone, special event, weather, or that no stored event on the segment exists. The vector of features is provided as input to the classification models for inference on the cause of the NRC.

Phase 2 : Classification models

Tree models where the target variable can take a finite set of values are called classification trees [74]. C4.5 is an algorithm used to build classification trees from a set of training data using the concept of information entropy [74]. The purpose is to split at each node with the feature having the highest normalized information gain. We employ such an algorithm in the training of the CT described in this paper.

We also developed a probabilistic model based on a Naive Bayesian classifier which gives useful predictions about the congestion. The aim of the Naive Bayesian classifier is to assign a target variable to one of a discrete set of categories based on its observable features.

$$P(Y|x) = \frac{P(Y)P(x|Y)}{P(x)}$$

Applied to our problem, translation of the a posteriori observable characteristics $x_1, x_2, ..., x_i$ into congestion component class I of $y_1, ..., y_j$ is computed by using Bayes rules :

$$P(I \in Y_j|x_1, ..., x_i) = \frac{P(I \in Y_j)P(x_1, ..., x_i|I \in Y_j)}{P(x_1, ..., x_i)}$$

The classifier is naive because it makes the strong assumption that the features are mutually conditionally independent ; that is, the conditional probability that I belongs to a particular class given the value of some feature is independent of the values of all other features. There is no statistically significant data for assessment of the more complicated causality between explanatory variables. Also, since the parameters of the NB model are estimated, probabilistic dependencies among features need contextual observations, the lack of ground-truth data prevents this research from fully modeling the realism of this transport-related phenomena. Despite this assumption, empirical studies demonstrate that it does not significantly compromise the accuracy of the prediction. This reduces the probability to :

$$P(I \in Y_j|x_1, ..., x_i) = \frac{P(I \in Y_j) \prod_{z=1}^{i} P(x_z|I \in Y_j)}{P(x_1, x_2..., x_i)}$$

I is typically assigned to the category with the greatest probability. The most likely j is chosen as follows :

$$j^* \in \arg\max P(I \in Y_j) \prod_{z=1}^{i} P(x_z|I \in Y_j) \tag{4.3}$$

and assigning I to class Y_{j*}.

Phase 3 : Cooperative process

If excessive travel time is detected on a segment, the scheme activates a cooperative process that shares the individual estimation made by the vehicle. We present in Fig. 4.2 the algorithm implemented on board of each vehicle and that is primarily for signalized arterials. We highlight the monitoring, aggregating, analysis and dissemination procedures. The purpose is to assess if the temporary induced traffic change related to an event can be mitigated in a short period or does the event represent a permanent change representing an NRC. We imple-

mented the methods described in [19], Basic Traffic Data Gathering algorithm, Local Traffic Evaluation Algorithm and Expanding the Evaluated Area algorithm. We add two types of messages : Extraordinary Event Request (RQ) and Extraordinary Event Response (RP). RQ is transmitted upstream via broadcast to all cars in its communication range, and allow to retain the event info locally on the segment for a minimum duration before propagating it to adjacent segments. In [10], it was shown that continuously high values of travel time along a segment that last at least four successive time intervals was the criterion used as evidence of a NRC event. This prevents false positive NRC detection. RP is the message used to send the NRC event to adjacent segments after the duration expires.

4.4 SIMULATION

The method provided in our study is applied to a heterogeneous network of both urban highways and signalized arterials [10]. The real-world traffic dataset of the Travel and Activity PAtterns Simulation (TAPAS) Cologne scenario [45] is considered a 'complex network' that mimics the real-life context of vehicle mobility. Heterogeneity exists on urban road networks where the structure of links varies substantially. The dataset comprise 700 000 individual car trips. Each line of the dataset contains the time, the vehicle identifier, its position and speed. Using SUMO, a microscopic traffic simulator for the simulation of urban mobility [72], we create extended scenarios mounted on top of the base scenario to model atypical traffic conditions such as weather, incident, workzone and special event. SUMO simulator needs two inputs : The Road Network of the city of Cologne is imported from the OpenStreetMap (OSM) database and the Traffic Demand is the dataset of car trips. The output of SUMO is the movement of vehicular nodes in a large urban network and data such as the acceleration, density, flows, gap between vehicles and other microscopic parameters at a vehicle level. From the simulation data collected by each vehicle, we extract features constituting an instance of the train dataset.

To simulate the Extended Scenarios of an Incident/Workzone, on the base scenario, we stop on a lane some vehicles for a specific amount of time. We vary the position on the edge and the duration. For the Extended scenario of bad weather, which lead to decreases in the vehicles' velocities and a more careful and defensive driver behaviour, we change the parameters of the car-following model in the simulator. For the special event scenario, to generate trips to a particular destination, we generate random departures and random routes. We use a Poisson process to generate random timings for trips. The rate parameter λ is the demand per second from different sources. To generate random routes, given trips are assigned to respective fastest routes according to their departure times and a given travel-time updating

DATA - *Vi* : Vehicle in the scenario, *oTT* : Observed travel time, *TTh* : Historical travel time, *CurrentTT(Vi)* : Travel time and local traffic evaluation, *TrajectoryTT(Vi)* : Travel time on edges of route stored in EdgesofRouteofV, *TrajectorySpeed(Vi)* : Speed of vehicles stored in ListeEdges, *TrajectoryDemand(Vi)* : Flows in ListeEdgesD, *TrajectoryGap(Vi)* : Gap distances in vehiclesG.

1: *Vi* broadcasts and recieves BEACON message from neighbors // MONITORING
2: Get current road segment of *Vi* and CurrentTT(*Vi*) on the segment // AGGREGATING
3: Update TrajectoryTT(*Vi*), TrajectorySpeed(*Vi*), TrajectoryDemand(*Vi*) and TrajectoryGap(*Vi*)
4: **if** $oTT > 1.8 * TTh$ **then** // ANALYSIS and DISSEMINATION
5: Calculate features
6: Create feature vector and Predict with BN
7: **if** StoredEvent == 0 **then**
8: *Vi* creates and backpropagates RQ
9: **else**
10: **if** Duration not reached **then**
11: Store RQ
12: **else**
13: **if** Duration reached or NRC is Incident or Workzone **then**
14: Backpropagate RP to adjacent road segments
15: **end if**
16: **end if**
17: **end if**
18: **end if**

Figure 4.2 Algorithm - Cooperative Process of VANET

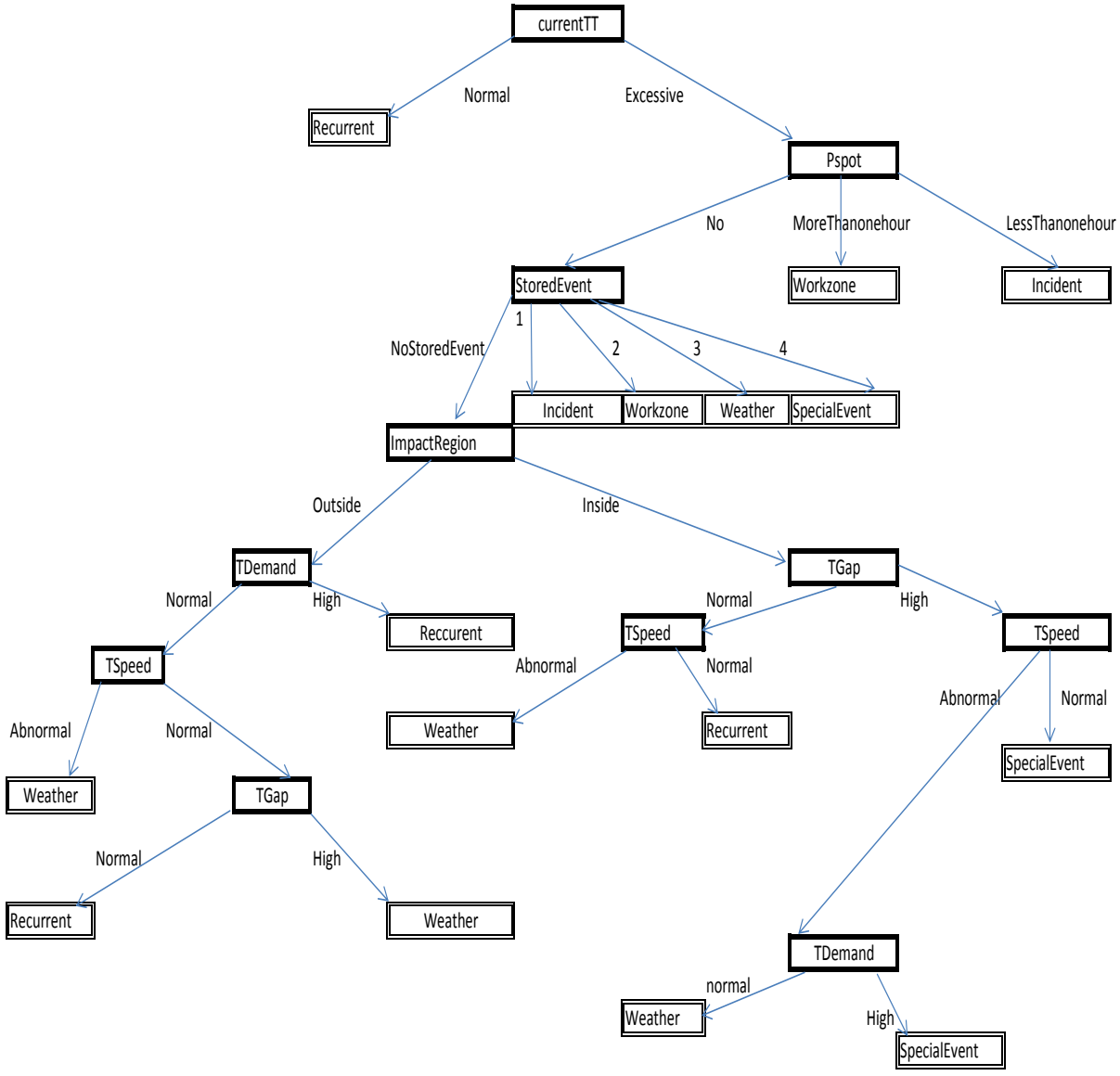

Figure 4.3 Classification Tree

interval by SUMO's traffic assignment model.

We construct a training dataset, a matrix with rows corresponding to samples and columns to features. The train dataset contains 591 instances. To obtain a realistic environment for the simulation of vehicular communications, we extract from the extended scenarios in SUMO, the vehicular traces that we will use in ns2 [76]. We assume that vehicles are equipped with a Global Positioning System (GPS) device for positioning, a transceiver for communication using Dedicated Short-Range Communications (DSRC), and an enriched digital road map containing information about the map, including the length of each road, number of lanes per road, SegmentID and the expected travel times on the segments. We use data forwarding techniques to pass information through the VANET such as geographical routing, and broadcast. For communication among all cars, we assume standard signal range of the 802.11p protocol, which is 300 meters.

4.4.1 Results

We demonstrate the robustness of our scheme by examining the performances of accuracy of classification, timing, and impact of NRC in an urban network. Firstly, we use Weka to generate a pruned classification tree [79]. Weka is a suite of machine learning software for data analysis and predictive modeling. The proposed CT is presented in Fig. 4.3. The accuracy of classification measures the predictive performance of the classifier and is determined by the percentage of the test dataset examples that are correctly classified. We performed 10 fold cross-validation on the training set and we got 87.63% of correctly classified instances. The value ranges of the splitting arcs are learned by the classifier and shown as nominal values on the arcs of the tree. The tree starts with CurrentTravelTime feature as the root node. CurrentTravelTime on a segment measures the travel time of a vehicle on each segment and compares it with TT_h of each segment. CT confirmed that when travel time on a segment is below its excessive treshold, the congestion is due to recurrent congestion. Otherwise, it is a NRC and if there is a problematic spot on the segment, the tree attributes the NRC cause to either an incident or a workzone. Then, the tree splits on the StoredEvent feature. If the vehicle is in a congestion due to a special event, its location is necessarily inside the impact region of the event. The tree splits on the internal non-leaf node labeled ImpactRegion. The leaf node SpecialEvent is on branches coming out of inside an impact region. TrajectoryDemand and TrajectoryGap diffrentiate data between a special event and a weather condition. Problems of small-data mainly revolve around high variance were outliers are present. More training cases are needed for the statistical inference to pick up the causality between explanatory features in order to make strong assumption.

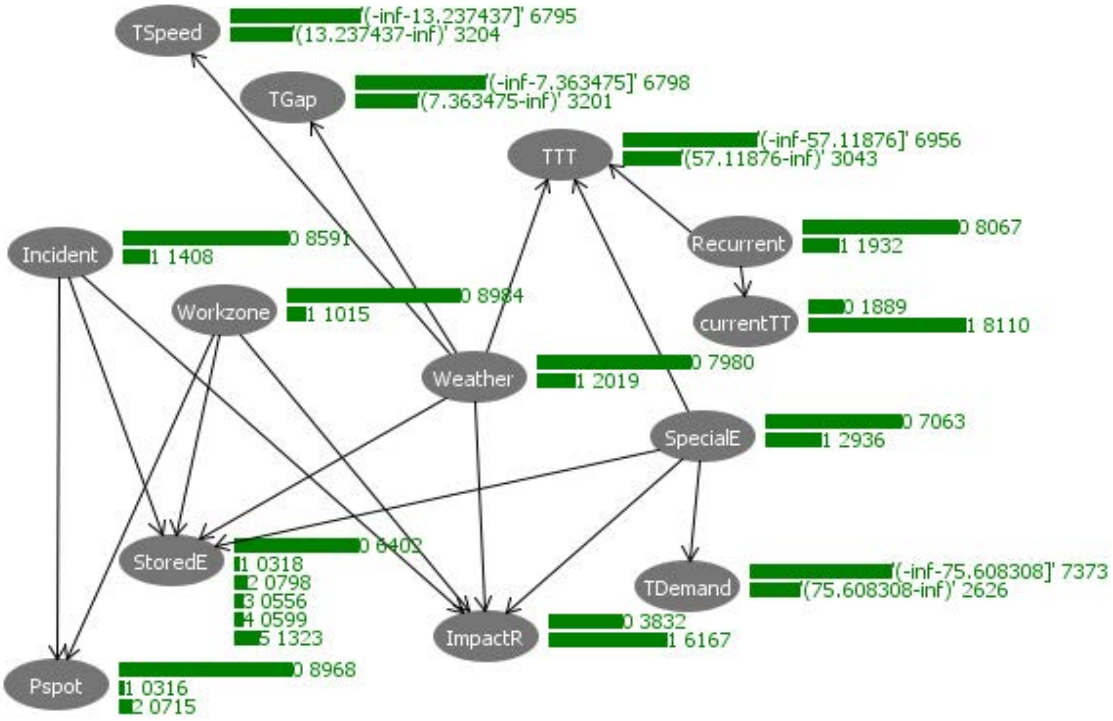

Figure 4.4 Bayes Network

BN is presented in Fig. 4.4. We performed 10 fold cross-validation on the training set to test the model. It's accuracy in terms of prediction error is 88.83% of correctly classified instances. Dependencies between a cause of congestion and its consequences are represented by arcs on the graph. All causes of congestion, except the recurrent congestion, have arcs going to ImpactRegion and SoredEvent features. This is because any NRC might occur inside an impact region as well as outside. Also, a StoredEvent feature will rule out all other causes of congestion if its value reports workzone, incident, weather or special event. We conducted a sensitivity analysis on the features of the model to note the importance of a feature for some partial classification. We removed one feature at a time and used the filtered training set for classification. We show in Fig. 4.5 the sensitivity of each feature on the accuracy of the CT and BN. We see that except for StoredEvent, the other features have the same performance.

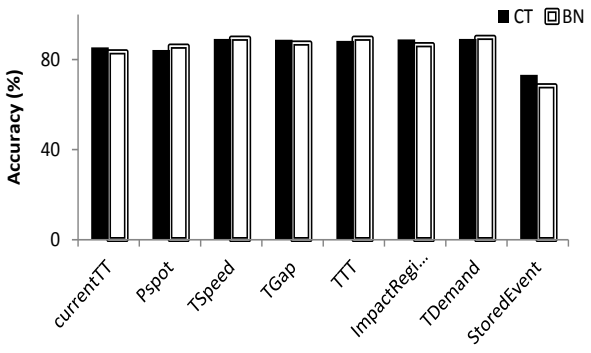

Figure 4.5 Sensitivity of the CT and NB models

A possible extension of the CT method described in our paper is Random forests. It's an ensemble learning method also used for classification. It's combining multiple models into ensembles to produce an ensemble for learning. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. As in CT, we use J48 to produce decision trees, then we produce slightly different decision trees by randomization. Weka's implementation yielded 89.51% accuracy of Random forest of 100 trees, each constructed while considering 4 random features.

Boosting is a fairly recent technique in supervised learning. AdaBoostM1 is a standard boosting scheme where diversity is created by focusing on where the existing model makes errors. Iteratively, new models are influenced by the performance of previously built models. Extra weight is given to instances that are misclassified to make a training set for producing the next model in the iteration. This encourages the new model to become an 'expert' for instances that were misclassified by all the earlier models. AdaBoostM1 is implemented in Weka. With 10 iterations, it yielded 89.17% of accuracy. In an urban road network context, the models were able to classify the cause of the congestion on both highways and signalized arterials because classification is based on the collection of features (problematic spot, currentTT, etc.) that are nonspecific to the type of the facility.

Secondly, vehicles in our scheme are able to monitor traffic and detect NRC during different times. In Fig. 4.6, we illustrate average travel time (TT1-TT4) of vehicles during incidents happening at different times (T1-T4 correspondingly) on the same busy road segment. Any time congestion is detected and the average travel time is above ExcessiveTT, we assess the number of vehicles reporting severe congestion compared to the number of vehicles on the segment. We highlight on Fig. 4.6 the time when congestion from incidents happens and denote it by TC. TC1 is related to the incident happening at T1 because the time of congestion differs from the time of the incident. The level of excessive travel time induced

by the incidents can be observed in the figure but it's only when congestion is detected at TC and values are higher than the threshold determined with the congestion factor that NRC is declared. Also, as in [10], continuously high values for at least four successive time intervals was the other criterion used as evidence of an event. The results indicate that on average, 88% of vehicles were able to detect the NRC and the percentage gradually increased to 95% in the next 5-15 minutes interval. We conclude that severe delay caused by NRC can be accurately detected any time it happens. To guarantee monitoring of road traffic at various level of granularity, a variable congestion factor could be considered interactively for management purposes. But for NRC detection, studies showed that a fixed congestion factor can accurately detect most NRC [10], as was the case in our experiments.

Finally, we demonstrate that vehicles are able via VANET to propagate the cause of NRC. For special events, we tested the accuracy of the impact region. We report the percentage of vehicles experiencing NRC due to a special event inside the impact region in contrast to all those experiencing the congestion from the special event, inside and outside the predefined impact region of the event. We monitor an impact region of only one road segment and we note that the size of the impact region has an initial impact on the detection rate, as seen in Fig. 4.7.

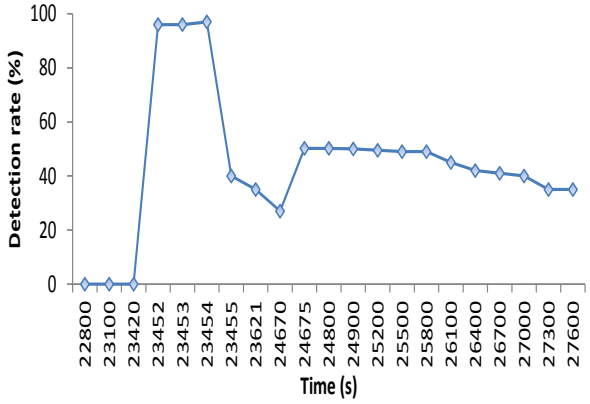

Figure 4.7 Accuracy of the impact region

Congestion from the special event starts around t= 22100s but it's only at t=23422s, that it becomes excessive. Vehicles inside the impact region are the first to detect the congestion and its cause and the detection rate is very high. After, the detection gradually decreases to 27.58% and increases again to 50.24% at t=24670s. This behavior is due to the cooperative process of our method. Communication between vehicles on the same segment has to happen for a certain duration before propagation of the event to first order adjacent segments can be done. During this period, more vehicles outside the impact region are experiencing severe

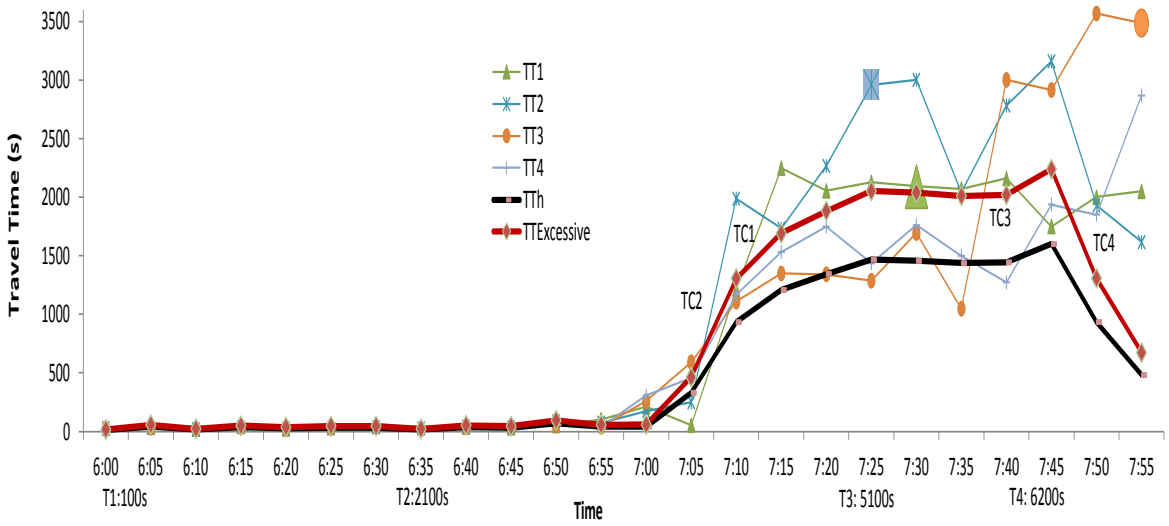

Figure 4.6 Average Travel Time on a signalized arterial during incidents happening at times T1 to T4

congestion from the special event but cannot accurately assess the cause, for this reason, detection rate decreased. Then, detection rate increased when more vehicles outside the region became aware of the event after communication between segments via VANET. The algorithm stops after the first-order adjacent segments ; consequently, detection rate cannot get any higher. The scheme does not evaluate the spatial extent of the NRC. It is out of the scope of this study. But the demonstration showed that the ImpactRegion feature can be used without any knowledge about the event because the region is reshaped via communication.

4.5 CONCLUSION

The duration, timing and location of non-recurrent congestion (NRC) in an urban network varies a lot making it difficult to monitor traffic in real time with conventional mechanisms. We have proposed a framework for the distributed classification of congestion into its components using VANET as an alternative cost-effective and flexible solution to guarantee better monitoring of road traffic on heterogeneous networks. The proposed framework aims to exchange traffic flow data and to embed reasoning machinery in vehicles to infer the cause of NRC.

We have obtained a predictive accuracy of 87.63% for the classification tree (CT), 88.83% for the Bayesian network (BN), 89.51% for Random forest (RF) and 89.17% for the boosting technique trained on synthetic data extended from the real case study of the Cologne scenario. In the future, we note that more sophisticated methods can be employed in the

cooperative process, such as a voting process, a likelihood evaluation or a model of the value of information. Also, data of connected vehicle operations in real-world conditions, such as Ann Arbor Automated Vehicle Operational Test, can be used as a test environment and provide real-world training dataset in occurance of different NRC scenarios [80].

CHAPTER 5 ARTICLE 2 : COOPERATIVE EVALUATION OF THE CAUSE OF URBAN TRAFFIC CONGESTION VIA CONNECTED VEHICLES

Ranwa Al Mallah, Alejandro Quintero, and Bilal Farooq
submitted to IEEE Transactions on Intelligent Transportation Systems

Abstract

We developed a distributed data mining based methodology to elaborate a decision concerning the cause of traffic congestion on a road network via emerging connected vehicle (CV) technologies. Our aim is to obtain deeper real-time insights of traffic conditions using decentralized cooperation between individual vehicles. We observe the complex phenomena through the interactions between vehicles exchanging messages via Vehicle to Vehicle (V2V) communication. Results are based on real-time data from vehicles experiencing traffic congestion on the simulation generated scenarios extended from the real-world traffic Travel and Activity PAtterns Simulation (TAPAS) Cologne scenario. We extract from the traces a dataset and evaluate a Voting Procedure (VP), Belief Functions (BF), a Data Association Technique (DAT) and an adapted β-DAT. Methods are tested and compared using a microscopic urban mobility simulator, SUMO and a network simulator, ns-2, for the simulation of communication between CVs. Compared to the Back-Propagation algorithm (BP) extensively used in the past literature, our performance evaluation shows that the proposed methods enhance the estimation of the cause of congestion by 48% for the proposed VP, 58% for the BF, 71% for the DAT and 70% for β-DAT. The methods also enhance detection time from 7.09% to 10.3%, and β-DAT outperforms BP by approximately 1.25% less false alarms triggered by the network, which can be significant in the context of real-time decision making. We show that a market penetration rate between 63% and 75% is enough to obtain the full benefits of V2V communications technology and ensure satisfactory performance.

5.1 Introduction

With the increasing number of vehicles and limited road network expansion, the urban traffic congestion is growing at an alarming rate. Urban environment exerts a profound influence over traffic, such as facilities and activities, weather, legal, social involvement and recurring incidents. Due to the high complexity and uncertainty of contemporary transportation sys-

tems, traditional traffic data collection and estimation tools fail to capture the dynamics in detail and in real time. Vehicular Ad hoc NETworks (VANET) known as connected vehicles (CVs), are key players' in the future self-organizing traffic information systems [81]. With the progress in information and communication technologies, CVs data collection and dissemination aims at building an intelligent public transportation system based on real-time information. For traffic management, the future will be in cooperative systems as they can benefit from the information collected from the mobile wireless vehicular ad hoc network.

A large body of work has already focused on the congestion detection problem using CVs [19] [82] [24]. Recently, a framework was proposed to further characterise the congestion detected. In [5], they tackled the problem of classification of congestion into its components in urban traffic. In fact, congestion can be classified as recurrent or non-recurrent. Recurrent congestion refers to congestion that happens every day on a regular basis [25]. Non-recurrent congestion in an urban network is mainly caused by incidents, work-zones, special events, adverse weather and bottlenecks [10]. In [83], if a vehicle detects congestion, it is able to predict with high accuracy its cause based on macroscopic and microscopic traffic variables that the vehicle collected along its trajectory.

Several unresolved problems exist for CVs-based congestion classification on urban networks. Firstly, each vehicle classifies individually the cause of congestion based on data from vehicles it encountered along its route. When congestion occurs, the vehicle tries to estimate the cause based on its experience. The assessment and classification are done locally at a vehicle level. If one vehicle sends a false alarms, it spreads uncertainty among vehicles and this in turn causes more congestion. The side effects of false alarms on the congestion level are a serious challenge because sending false information disrupts the proper network operation. This behaviour is a threat to the traffic network and in terms of security, it is comparable to the *simulation of multiple entities* or *sybil* attack in which an attacker uses different identities at the same time to send false information or simulate a false congestion. Besides, Sybil attack are very hard to detect, particularly in such highly dynamic environment like VANET. This makes exploring the cause of congestion at a vehicle level a partial limited solution since occasionally, honest vehicles in this scheme behave as malicious users without their intent but because of the scheme.

Secondly, in the schemes proposed in the literature each vehicle classifies individually the type of event based on its personal information [40] [42], but because traffic is multifaceted, we warn that the vehicle has partial knowledge about the road condition, it knows to some degree the traffic condition surrounding it. This decreases the estimation accuracy of the real cause of congestion. Schemes should be implemented to obtain deeper insight on the cause of

traffic congestion using cooperation between individual vehicles. This will increase estimation accuracy because inaccurately estimating the cause of congestion misleads other vehicles as well as traffic controllers and leads to devastating consequences. It has similar consequences on the network as false alarms.

Thirdly, in urban networks vehicles are repeatedly faced with situations where they encounter congestion. Vehicles will have to repeatedly determine its cause based on the variables they collected. In addition to connectivity, they should be intelligent enough to learn from their experiences. Currently, there is no mechanism that is able to extract valuable knowledge from the situations experienced by the vehicles. Every situation should be a suite of instances learned for better decision making because the monitoring currently done by the proposed schemes does not allow for summarizing valuable knowledge. Since congestion in an urban network is mainly caused by incidents, work-zones, special events, adverse weather, or recurrent congestion [10], for a given situation, the classification algorithm proposed in the literature returns the cause with the greatest probability as the most likely cause [83]. In fact, a classification algorithm is a procedure for selecting a hypothesis from a set of alternatives that best fits a set of observation. In other words, although the classifier is able to compute the probability that a traffic situation belongs to a particular cause given the value of some features, the algorithm returns only the cause with the highest probability. For example, if the classifier computed a probability of 0.32 for the cause of congestion being an incident and 0.31 for the cause of congestion being due to a work-zone, then the most likely cause of congestion selected by the classification algorithm is an incident. We propose that we can make use of each probability computed by the classifier to extract more knowledge.

The goal here is to obtain deeper insights on the cause of traffic congestion using cooperation between individual vehicles. Since urban traffic is essentially unstable, chaotic, and unpredictable, individual vehicle assessment is not enough, the next step is to elaborate a decision collectively. This would result in a more precise, efficient, and reliable view of the traffic condition by observing the complex phenomena from the interactions between vehicles. In the current state, if an event received by a vehicle is a false alarm, the algorithm will fuse the obtained information with others located on a same road segment and spread uncertainty among vehicles and this in turn causes more congestion. An evaluation process has to take place after data sensing and before data fusion. We add this layer to address the vulnerability of fusion algorithms and to lower the side effects of false alarms because the approaches proposed in the literature fail to process the data before fusion and present a security threat to the network. Furthermore, since information is a subject of interest to the vehicles in a given geographical area, the methods we propose elaborate a decision collectively on a given geographical area to obtain deeper insights of traffic condition before fusion is applied.

Also, we propose that each vehicle represent its uncertainty about the cause of congestion in a vector of probabilities associated to each of the possible causes of congestion before exchanging the vector with the vehicles on the road segment. We explore the collected vectors for learning purposes by building a dataset and extracting relationships via data mining techniques to develop useful patterns and information. Particularly, this data analysis is used to build models capable of machine learning.

Our mining methods consist of a voting procedure, belief functions and a data association technique for efficient inference on the cause of traffic congestion via CVs technology. We consider a realistic map configuration of the city of Cologne in the evaluation of our methods. Compared to the backpropagation (BP) technique proposed in the literature [83], the proposed techniques enhance the estimation accuracy by 48% for the proposed VP, 58% for the BF, 71% for the DAT, and 70% for the adapted data association technique (β-DAT). The methods also enhance the detection time by 10.3% for VP, 9.40% for BF, 9.45 for DAT and 7.09% for β-DAT. β-DAT outperforms BP by approximately 1.25% less false alarms triggered by the network. The methods also require only 63% penetration rate to obtain the full benefits of V2V. Knowing the root causes of congestion that are affecting their facilities will enable road authorities to make more informed decisions about how to best reroute traffic, change lane priorities and modify traffic light sequences. It may also assist the road authorities for better planning of road network expansion, as well as optimal road sign placement and speed limit setting.

The contributions of this paper are summarised as follows :

— Addition of an evaluation layer before fusion can take place in order to lower false alarms that are comparable to security threats on the traffic network.
— Implementation of a cooperation process to increase estimation accuracy because traffic is multifaceted and to conceal the fact that individually, vehicles have partial knowledge about the road condition.
— Generation of a dataset for association rules mining to extract more knowledge and implementation of the rules on board of the vehicles for analysis and evaluation of the cause of urban traffic congestion with short-range communication between vehicles, Vehicle-to-Vehicle (V2V), being the communication architecture for seamless decentralized exchange of information between the cooperating vehicles.
— Validation of the methods using a microscopic urban mobility simulator, SUMO [72] and a network simulator, ns-2 [76], for the simulation of communication between CVs.
— Evaluation of the influence of the penetration rate of CVs on the methods and detection of the necessary market penetration rate of V2V communications technologies on the

performance of the methods for transportation to obtain the full benefits of V2V communications.

This paper is organized as follows. A review of related studies is provided in Section 5.2. The methods are detailed in Section 5.3. In Section 5.4, we provide results, analysis and discussion. Finally, conclusion and future work are outlined in Section 5.5.

5.2 RELATED WORK

A traffic management system consists of a set of complementary phases, each of which plays a specific role in ensuring efficient monitoring and control of the traffic flow in the city [4]. In Fig. 5.1, the data sensing and gathering phase, heterogeneous road monitoring measures traffic parameters such as traffic volumes, speed, road segments' occupancy, etc. Subsequently, these data feeds are fused and aggregated to extract useful traffic information. This acquired knowledge from the processed data is used in the data exploitation phase to compute optimal routes for the vehicles, short-term traffic forecasts to reduce road traffic congestion, improve response time to incidents, and ensure a better travel experience for commuters. Finally, in the service delivery phase, the traffic management system delivers this knowledge to the end users.

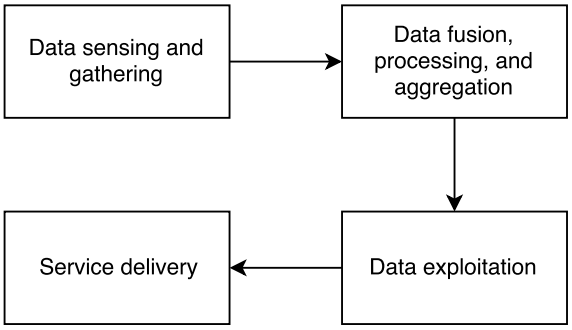

Figure 5.1 Different phases of data in a traffic management system

In the first phase, traffic data can be collected from fixed monitoring equipment, such as induction loops, sensors and CCTV cameras. However, in this scheme, discovering the dynamic properties of the traffic is a difficult task due to the sparseness of the deployed equipments [81]. On the other hand, monitoring can be done using mobile data sources such as GPS-based systems, floating car data, SMS, social data feeds, etc [4]. In this case, the challenge is the low penetration rates of the entities transmitting the data. In both cases, deploying highly sophisticated equipment to ensure the accurate estimation of traffic flows and timely

detection of events may not be the ideal solution, due to the limitation in financial resources to support dense deployment and the maintenance of such equipment, in addition to their lack of flexibility. In sum, the currently deployed technologies for road traffic surveillance still suffer from a lack of traffic parameter measurement accuracy to enable granular and timely monitoring of events that occur on the roads.

New technologies can be used to improve the accuracy, timeliness, and cost efficiency of data collection. In fact, researchers have been focusing their efforts on exploiting the advances in sensing, communication, and dynamic adaptive technologies to efficiently monitor the evolving critical road infrastructure [4]. The application of wireless technology to moving vehicles enables the creation of vehicular ad hoc networks, also called VANET. Connected vehicles of the VANET are scalable enough to enable better control of the traffic flow and enhance management of large cities' road networks [32]. Many studies have been proposed in the literature to illustrate that CVs improve the accuracy of the acquired real-time traffic information [43]. In our study, we use CVs to monitor the traffic parameters experienced by a vehicle along its trajectory.

Subsequently, the vehicles periodically exchange the data collected with the vehicles in their surroundings. Algorithms based on V2V communications are very different from algorithms developed in vehicle to infrastructure (V2I) communication applications [43] [4]. In the latter study, a centralized module combines collected data and disseminates global information. This present work concerns V2V communication mode where vehicles do not use any centralized access point to build their own information assembly. Particularly, different strategies have been proposed in the literature [84] [85] concerning the problem of information dissemination, i.e., proposing a strategy to exchange information between the CVs. Indeed, when the road traffic is high, the bandwidth is limited and the number of message exchanges would have to be reduced. In our work, we can use any of the information dissemination strategies proposed in the literature to exchange information between vehicles because our aim is rather to explore the exchanged data by using data mining techniques in order to elaborate a decision collectively regarding the traffic condition experienced by the vehicles.

In the second phase, data fusion algorithms take the data collected from the information dissemination process and use it to improve : the reliability of a judgment by the contribution of redundant information ; or the interpretation ability by the provision of complementary information. Particularly, a large portion of literature has been proposed for the distributed data fusion for uncertain reasoning in ad hoc and dynamical networks [39] [40] [41] [42]. In [39], they introduced belief functions to combine and fuse data in vehicle for the management of uncertainties about events in vehicular networks. The theory of belief functions is a gene-

ralization of the Bayesian probability theory. Belief functions combine degrees of confidence about events reported in exchanged messages.

Specifically, concerning spatio-temporal events such as traffic congestion, in [40], belief regarding the presence of an event on a geographical point is obtained by : discounting [41] neighbouring information according to their distance from the point ; then combining the obtained information [42]. The authors propose to use the cautious combination rule [44] to fuse information located on a same road segment. In [42], strategies to fuse acquired information consider message aging of local events. They extend the work in [41] by developing new methods based on the notion of update and by proposing away to automatically compute the message aging (by discounting or reinforcing) using historical data. In [45], unlike the model in [40], was the choice of the event dissemination strategy considered. Each vehicle sends new events or repeats received one. A choice has been undertaken to keep combinations of messages in each vehicle.

There are two major drawbacks to these approaches. Firstly, if the event received by the fusion algorithm is a false alarm, the algorithm will fuse the obtained information with others located on a same road segment and spread uncertainty among vehicles and this in turn causes more congestion. The methods we propose elaborate a decision collectively on a given geographical area to obtain deeper insight of traffic condition before fusion is applied. We add this layer to cope with the vulnerability of fusion algorithms and to lower the side effects of false alarms. Secondly, the approaches fail to process the data before fusion. The methods manage uncertainties by combining degrees of confidence about events reported in exchanged messages based on attributes such as the geographic distance of the event from the receiver and message aging of the event. Unlike the approaches proposed in the literature, we deal with uncertainties by the accurate evaluation of the cause of the congestion by cooperation between vehicles on the road segment before the management methods can take place. Since information is a subject of interest to vehicles in a given geographical area, an evaluation process has to take place in this phase before data fusion and aggregation. Our work can then use any of the management strategies proposed by all these works to fuse information.

For validation, in [39] the methods were tested and compared using a Matlab simulator where roads are divided into segments and one event is considered per segment. In [45], the model was implemented and tested using Hong-Ta Corporation (HTC), an application using smartphones. The application proposed required driver assistance to send the events and the authors proposed that camera or sensors might be installed in vehicles in such a way to automatically detect events. In [41], authors propose to divide map traffic lanes into

small rectangular areas named cells. The map is composed of horizontal and vertical two-way streets in particular a traffic lane is composed of *NbSimCells* cells depending on the type of event. A mechanism allows smoothing results by considering neighbouring influence. Results depend on method cells size and influence mechanism rate σ. The automatic computation of the method cells sizes and the values of σ are still underdevelopment ; currently they are set manually. The authors propose a possible solution to use historical knowledge to study these parameters or to choose a small method cell length and have a short time step if bandwidth and databases spaces permit. For example, supposing that traffic jam takes place starting from ten successive vehicles driving very slowly ; then the method cell length is equal to ten times the average length of a vehicle. Authors suggest that sensors might be used to detect events in order to create messages automatically, without driver assistance. In [42], they propose an influence mechanism to predict overall road situation based on the fact that traffic jams evolve in the reverse direction of traffic lanes and disappear in the same direction of roads. For each cell of length 67 m, on which a vehicle has information about the presence or the absence of a traffic jam event, m is the result of the fusion f mass functions of all stored messages concerning this event, the influence of m is the discounted mass function where 1 - β is the discounting rate. The influence mechanism consists for each cell c in combining conjunctively : obtained influences on cell c and the result of the combination of mass functions of all created or received messages. If m informs that a traffic jam is present on the cell c, the vehicle generates influences on following cells and stop this operation when arriving to a slowing down event exit. The mechanism requires a slowing down event exist to be generated by the vehicles of the cells. A slowing down event can be : related to map infrastructures and always present on the map (the map is known by all vehicles) as a roundabout ; or an event on the road known in vehicle database like an accident. Since their influence mechanism predicts the transfer of traffic, different causes of congestion require different types of management for the mechanism to work properly and not generate false influences. They stated that unlike traffic incidents, the spatiality of fog blankets does not depend on maps and to manage this event, roads are divided into cells without taking into account traffic directions. In other words, if a fog blankets event is present on one side of a traffic lane, it is also certainly present on the opposite side. The simulator the authors used is a research tool ; they suggested that coupling their method with an ad hoc network simulator will be a real added value for validation. The authors recommended TraNS [44] simulators : it combines SUMO [72] (mobility simulator) and NS-2 [76] (network simulator) simulators.

In our work, we use SUMO and NS-2 for the simulations of our scenarios. Moreover, unlike the previous literature, we consider a realistic map configuration of the city of Cologne in

the evaluation of our methods. Also, we simulate experiments on the real-world traffic Travel and Activity PAtterns Simulation (TAPAS) Cologne scenario. To the best of our knowledge, this is the first attempt to assess in the most realistic way, the reality concerning traffic jam events on the road for the sake of accurate estimation of the cause of congestion via CVs.

Finally, the reality concerning traffic congestion can be better assessed if we look into the causes of the traffic congestion. In [83] vehicles detect excessive congestion on a road segment and are able to estimate with high accuracy its cause based on a classification algorithm implemented on board of each vehicle. A classification algorithm is a procedure for selecting a hypothesis from a set of alternatives that best fits a set of observation. Since congestion in an urban network is mainly caused by incidents, work-zones, special events, adverse weather, or recurrent congestion, the classification algorithm returns the cause with the greatest probability, the most likely cause. In other words, although the classifier is able to compute the probability that a traffic situation belongs to a particular cause given the value of some features, the algorithm returns the cause with the highest probability. For example, if the classifier computed a probability of 0.32 for the cause of congestion being an incident and 0.31 for the cause of congestion being due to a work-zone, then the most likely cause of congestion selected by the classification algorithm is an incident. We discovered that we can make use of the other probabilities computed by the classifier to extract more knowledge. In fact, we propose that each vehicle represent its uncertainty about the cause of congestion in a vector of probabilities associated to each of the possible causes of congestion before exchanging the data with the vehicles on the road segment. Vehicles in the surrounding collect and evaluate the data before fusion can take place. We explore the collected data for learning purposes by building a dataset and extracting relationships via data mining techniques to elaborate a decision on the current traffic condition on the road segment. In fact, data mining techniques such as clustering, association, classification, have been applied in VANET to extract useful patterns and information [86]. Particularly, association rules mining is useful for data analysis, and is used to build models capable of machine learning. By simulating numerous scenarios, on different road segments, we can learn from the macroscopic and microscopic parameters the vehicles collected. The mechanisms for the exchange and management of the events are beyond the scope of this paper.

These systems cannot be deployed in the near future, as one has to wait until the necessary market penetration of V2V communications technologies has been reached. The fact that few vehicles are equipped with transceivers leads to network fragmentation [4]. As there is no other exchange of traffic information, its performance strongly depends on the penetration rate of participating vehicles. However, the question that arises in such a case is what percentage of penetration is enough to obtain the full benefits of the methods. We study the impact of the

market penetration rate on the performance of the methods.

Unlike previous works, we seek to improve the vehicle's estimation of the cause of congestion and reduce false alarms by cooperative methods infused with knowledge about the others evaluation specifically in the event of traffic congestion. Also, vehicles in our methods make use of each probability computed by the classifier to extract more knowledge. We prove that this data mining also reduces false alarms. These methods will leverage some properties of the road network such as the spatiotemporal correlation for efficient estimation of the cause of traffic congestion. The result is a voting procedure, belief functions and a data association technique for efficient evaluation of the cause of urban traffic congestion via CVs.

5.3 DATA MINING METHODS

Congestion in an urban network is mainly caused by incidents, work-zones, special events, adverse weather, or recurrent congestion. Vehicles are equipped with a method to detect excessive congestion in an urban network and a classification algorithm able to attribute a possible cause to it, as in [83]. The classification algorithm returns the cause with the greatest probability, the most likely cause. We make use of the other probabilities computed by the classifier to extract more knowledge. In fact, we propose that each vehicle represent its uncertainty about the cause of congestion in a vector of probabilities associated to each of the possible causes of congestion. In particular, the vector of probabilities exchanged between the CVs should have this form :

C = [P$incident$, P$workzone$, P$weather$, P$specialevent$, P$recurrent$]

After proper representation of vector C, different methods to elaborate a decision concerning the cause of urban congestion on the segment are presented in this paper. The environment is without infrastructure, i.e., there is no centralized access point that collects and disseminates global data on the road segment. Each vehicle has its own representation based on macroscopic and microscopic traffic variables the vehicle collected along its trajectory. When vehicles experience excessive congestion, they exchange via broadcast the representation stored in vehicle as illustrated in Fig. 5.2. Each vehicle also has its own decision module for the cooperative evaluation of the cause of congestion experienced. The decision module contains one of the Voting Procedure (VP), Belief Functions (BF) and Data Association Technique (DAT) methods described in the following sections.

A. Voting procedure

The voting procedure (VP) starts when a vehicle experiences excessive congestion and wants to evaluate the cause of the urban traffic congestion on the road segment. From [83], the

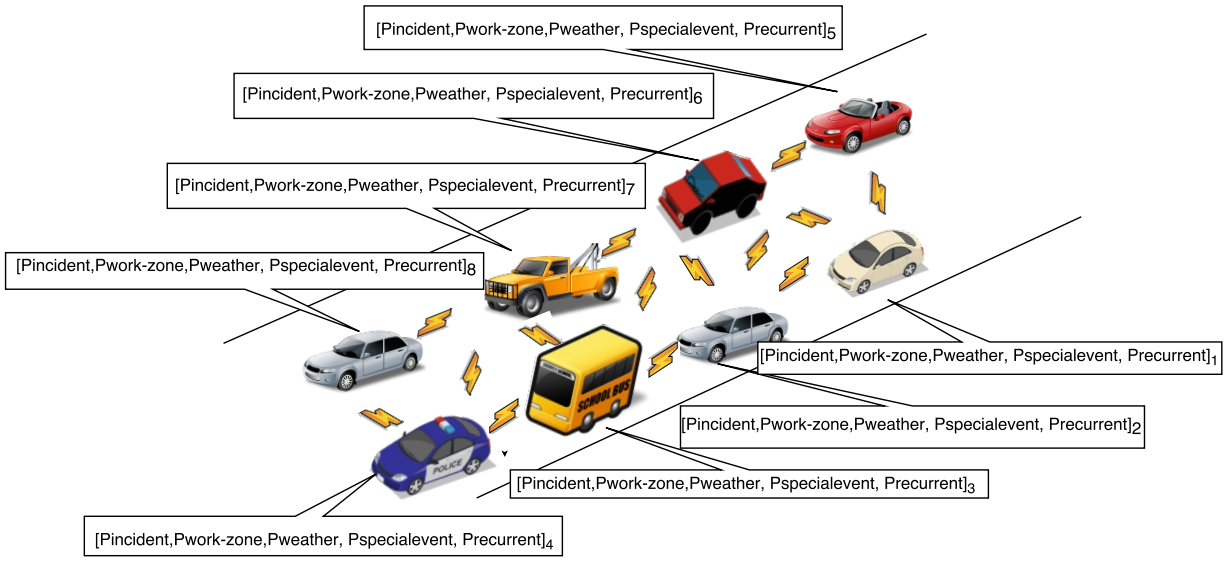

Figure 5.2 Vehicles exchanging via geographic routing information about the cause of congestion

probabilistic classification model on board of each vehicle predicts the cause of congestion but we extract the rest of the probabilities and the result is in the form of the probabilities vector presented above, with one cause of congestion having the highest probability. The vehicle broadcasts the probabilities vector in the vicinity for efficient evaluation of the cause of urban traffic congestion with short-range communication between vehicles, Vehicle-to-Vehicle (V2V), being the communication architecture for seamless decentralized exchange of information between the cooperating vehicles. Vehicles on the road segment collect the messages received and the decision module on board of each vehicle computes the counts for each cause. The cause having the highest count is highlighted by this voting procedure as being the cause of congestion on the road segment.

The VP is an improvement of the algorithm for backpropagation presented in [83]. Nonetheless, if vehicles vote, some vehicles cannot quantify their ignorance on the presence or the absence of congestion. In other words, their vote is not an accurate representation of information. Belief functions avoid this problem because partial or total ignorance can be represented.

B. Belief functions

The aim of this belief functions method (BF) is to improve the level of knowledge and thus enhance the prediction accuracy of the cause of congestion experienced by vehicles on the road segment. When reasoning with epistemic uncertainty, due to lack of knowledge (partial

knowledge) and uncertain information, uncertainty can be reduced. Probability theory can be used to represent epistemic uncertainty. In this case, probabilities are subjective, interpreted as degrees of belief. The main objection against the use of probability theory as a model of epistemic uncertainty is its inability to represent ignorance. The principle of Indifference states that in the absence of information about some quantity X, we should assign equal probability to any possible value of X. Also, probability theory is not a plausible model of how people make decisions based on weak information. Set-membership approach is another framework that can be used to represent epistemic uncertainty. Partial knowledge about some variable X is described by a set of possible values E (constraint). The advantage is that it is computationally simpler than the probabilistic approach in many cases because it's an interval analysis. But the drawback is that there is no way to express doubt making it a conservative approach. The theory of belief functions [39] extends both the Set-membership approach and Probability Theory. The theory includes extensions of probabilistic notions (conditioning, marginalization) and set-theoretic notions (intersection, union, inclusion, etc.).

Let $\Omega=\{\omega_1, \omega_2, ..., \omega_k\}$ denotes a finite set containing all the possible answers to a given question Q of interest ; Ω being called the *frame of discernment*. Information given by different sources regarding the answer to question Q can be represented by a *basic belief assignment* (BBA), also called a mass function, denoted by m. It is defined from 2^Ω (the set of all possible subsets of Ω) to [0,1] such that the sum of all the masses is equal to 1 :

$$\sum_{A\subseteq\Omega} m(A) = 1 \tag{5.1}$$

A mass $m(A)$ represents the belief supporting A, where A is a subset of Ω. It is the mass allocated to the hypothesis : the answer to question Q belongs to the subset A of Ω. Each subset A of Ω such that $m(A)> 0$ is called a *focal element* of m. The theory of belief functions allows the allocation of belief to subsets of Ω with no influence on the singletons, contrary to the probability theory [40]. Note that due to a lack of information, the part of belief cannot always be given to a singleton. The mass $m(\Omega)$ represents the degree of ignorance of the source which has provided the information m. The mass on the empty set $m(\emptyset)$ represents the conflict.

In case two vehicles express their beliefs over the frame, to combine the independent sets of probability mass assignments, $m1\cap2$, quantified by m_1 and m_2 and expressed on Ω, Dempster's rule of combination is the appropriate fusion operator. This rule derives common shared belief between multiple sources and ignores all the conflicting (non-shared) belief through a normalization factor. Specifically, the combination is calculated in the following

manner :

$$m1 \cap 2(A) = \frac{1}{1-K} \sum_{B \cap C = A} m_1(B) m_2(C), \forall A \subseteq \Omega \tag{5.2}$$

where

$$K = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \tag{5.3}$$

K is a measure of the amount of conflict between the two mass sets. With this combination, masses are transferred to focal elements intersections. The TBM postulates that uncertain reasoning and decision making are two fundamentally different operations occurring at two different levels : Uncertain reasoning is performed at the credal level using the formalism of belief functions. Decision making is performed at the pignistic level, after the mass function on Ω has been transformed into a probability measure. The pignistic transformation $BetP$ transforms a normalized mass function m into a probability measure as follows :

$$BetP(\{\omega\}) = \sum_{A \subseteq \Omega, \omega \in A} \frac{m(A)}{\mid A \mid (1 - m(\emptyset))}, \forall \omega \in \Omega \tag{5.4}$$

Applied to our problem, a mass function m is held by each vehicle and is defined on the frame of discernment Ω={Incident, Workzone, Weather, SpecialEvent, Recurrent}. Each vehicle assigns a mass on any of the singletons and another on a subset containing the singleton. This strategy is possible because of the classification model implemented on board of each vehicle. In fact, the probabilistic Bayesian network in [83] infers on the cause of congestion based on the macroscopic and microscopic traffic variables the vehicle collected along its trajectory. The model is not only able to return the cause having the greatest probability, but it can also return the 2-item subset containing that singleton and its probability. In other words, if the most likely cause of congestion computed by a vehicle is an incident, then the model can also inform that the second best possible cause it predicted is a weather condition. We consider this 2-items subset, {Incident, Weather} and transfer the Bayesian probability of the subset to a mass function. The subset containing the singleton represents added knowledge about the sensed traffic condition. We limit the strategy to 2-items because results showed very little improvement in the accuracy of prediction when more items are considered.

Many studies had proved and validated the accuracy of the theory of belief functions for the distributed data fusion for uncertain reasoning in vehicular ad hoc networks [39] [40] [41]

[42]. We specifically applied this theory to the uncertain reasoning about the cause of traffic congestion experienced on a road segment. We isolated the analysis in order to get more insight on the reasoning. In the next section, we propose to collect the fusion results for the sake of learning. We believe that knowledge can be acquired from the fused data that the vehicles exchange in a presence of a particular road condition. In fact, data mining techniques such as association rules mining have been applied in VANET to extract useful patterns and information. It is useful for data analysis, and is used to build models capable of machine learning.

C. Association rules mining method

A data association technique (DAT) is presented in this section to obtain more information about the cause of the congestion. Since each vehicle's assessment is communicated to vehicles in the vicinity, we collect the vector of probabilities exchanged by the vehicles in many scenarios to build a dataset. We extract the general association rules from the messages exchanged regarding the cause of the congestion. We analyse the messages for frequent patterns in order to identify the relationships for rule generation. Consequently, data association between messages exchanged by CVs will help analyze the road condition. Our aim is to improve the level of knowledge from exchanged messages for efficient evaluation of the cause of congestion. Let *Ims* be the set of all possible items *Ims* = { i_1, i_2, i_3, i_4, i_5}. Applied to our problem, i_1= Incident (I), i_2=Work-zone (Wo), i_3= Weather (We), $_4$= Special Event (SE), i_5= Recurrent (Re).

Also, let t_i be a subset of items also called an itemset, T is the set of all transactions T={ t_1, t_2, t_3... t_N} and N being the total number of transactions in the dataset. Let support indicate how frequent items appear in the dataset i.e. the number of transactions that contain a particular itemset X.

$$\sigma(X) = |\{t_i | X \subseteq t_i, t_i \in T\}| \tag{5.5}$$

Association rules are *if/then* relationships that help uncover seemingly unrelated data in a relational database. There are two parts, the antecedent is the *if*, found in the data and the consequent is the *then*. If X and Y are disjoint sets $X \cap Y = \emptyset$, for the rule $X \rightarrow Y$, we use the two criteria, support and confidence to identify relationships.

— support of X→ Y :

$$s(X \rightarrow Y) = \sigma(X \cup Y)/T \tag{5.6}$$

— confidence of X→ Y : number of times if/then statement have been found true.

$$c(X \rightarrow Y) = \sigma(X \cup Y)/\sigma(X) \tag{5.7}$$

It is how frequently items in Y appear in transactions that contain X. It measures the reliability of the inference made by a rule. It also provides an estimate of the conditional probability P(Y|X). We note that inference made by an association rule does not imply causality. It suggests strong co-occurrence relationship between items in antecedent and consequent of a rule. Causality requires knowledge about the causal and effect attributes in the data.

The problem can be stated as follows : Given a dataset, find all the rules having support >= *minsup* and confidence >= *mincon*, *minsup* and *mincon* are thresholds derived from the dataset. Support of rule X→ Y depends on support of its corresponding itemset $X \cup Y$, frequency of $X \cup Y$ in T. Fig. 5.3 and Fig. 5.4 show examples of how transactions can be collected from different scenarios to build a dataset for learning. If a road segment is congested due to a special event occurring in the surrounding area as in Fig. 5.3, firstly, each vehicle computes a probabilities vector. Then, each vehicle constructs a transaction composed of two items. If the probabilities vector [P*incident*, P*workzone*, P*weather*, P*specialevent*, P*recurrent*] of vehicle A is [0.15, 0.12, 0.23, 0.3, 0.2], this means that the cause having the highest probability 0.3 corresponds to a special event (SE). The second probable cause of congestion as evaluated by vehicle A is attributed to a weather condition (We) with a probability of 0.23. Vehicle A creates a transaction T_A ordered as follows SE, We, with items SE and We being categorical elements. In Fig. 5.4 for example, we collect transactions from a scenario simulating an incident.

A common strategy adopted by many association rule mining algorithms is to decompose the problem into two major subtasks.

Frequent itemset generation : We analyse data for frequent patterns by determining support count in the transactions for each candidate itemset. The objective is to find all the itemsets that satisfy the minimum support threshold. These itemsets are called frequent itemsets. According to the Apriori principle theorem, if an itemset is frequent, then all of its subsets must also be frequent. Support-based pruning is done first to trim the exponential search space because of the property that the support for an itemset never exceeds the support of its subsets, it's the anti-monotone property of the support measure. We adapt the pseudocode for the frequent itemset generation part of the Apriori algorithm as follows : - Generate the list of all possible itemsets. With a dataset that contains k items, it can possibly generate up to 2^k-1 frequent itemsets, excluding the null set. If k is very large, the search space of itemsets

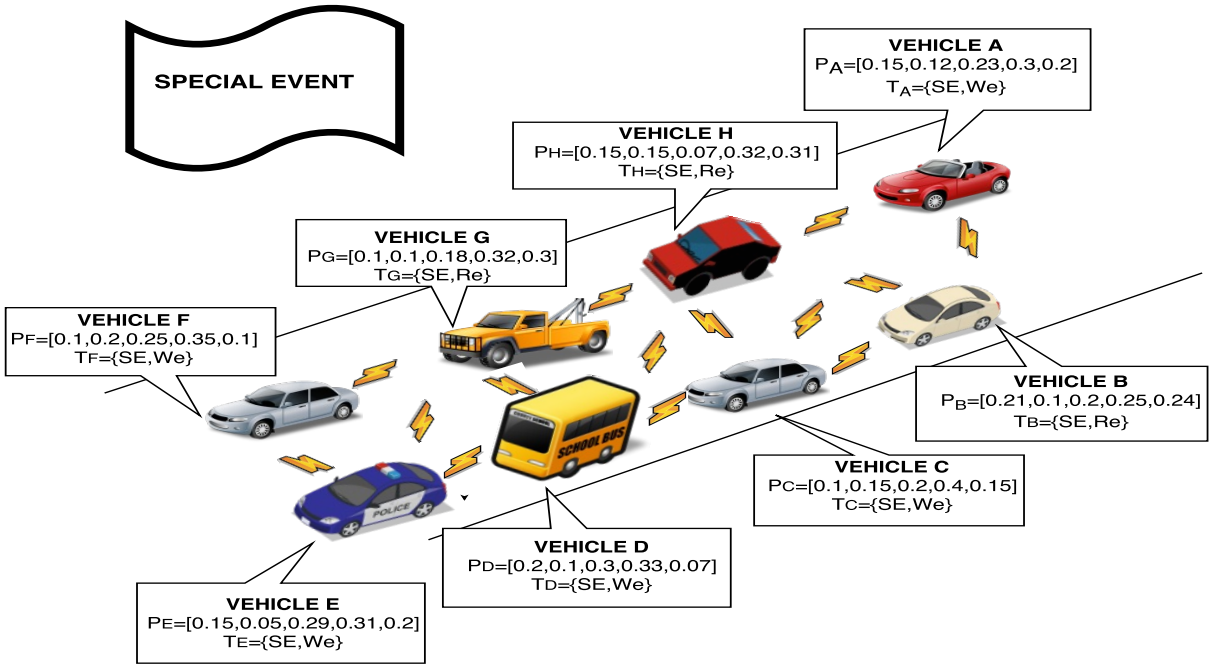

Figure 5.3 Transactions created by vehicles on a congested road segment due to a Special event

that need to be explored is exponentially large. - Determine support count for every candidate itemset. - Compare each candidate against every transaction. - If candidate is contained in a transaction, its support count will be incremented. In an initial iteration, the algorithm makes a single pass over the dataset to determine the support of itemsets containing one item, candidate 1-itemsets. Upon completion of this step, the set of all frequent 1-itemsets, F_1, will be known.

F_k= {i| i ∈ Ims ∧ σ({i}) >= N*minsup}

By replacing σ({i})>=threshold with max(σ({i})), we find all maximum 1-itemsets as per the voting procedure described in the previous section.

Next, the algorithm will iteratively generate new candidate k-itemsets using the maximum (k-1) itemsets found in the previous iteration. Candidate 2-itemsets are generated using only the frequent 1-itemsets because the Apriori principle ensures that all supersets of the infrequent 1-itemsets must be infrequent. After counting the support of the 2-itemsets candidates, the algorithm eliminates all candidates itemsets whose support counts are less than a threshold, by replacing the extraction of frequent k-itemsets by the extraction of the maximum support k-itemset :

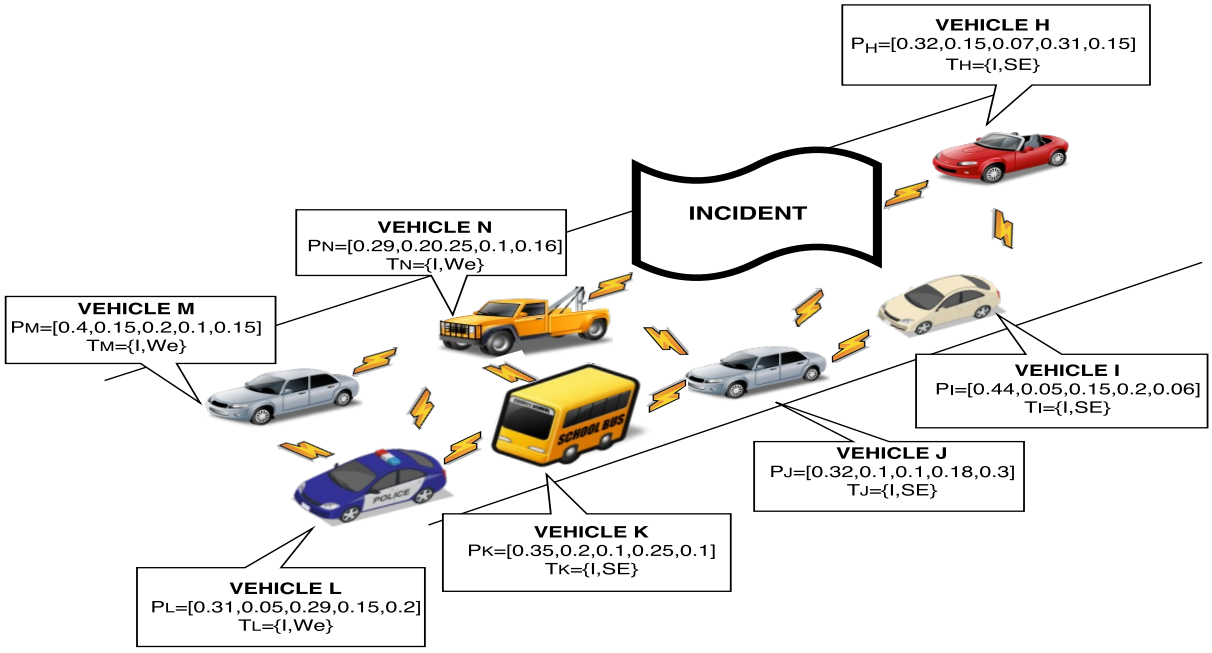

Figure 5.4 Transactions created by vehicles on a congested road segment due to an Incident

$F_k = \{c |\ c \in C_k \wedge \max(\sigma(\{c\}))\ \}$

The algorithm terminates when there are no new itemsets to be generated, i.e. $F_k = \emptyset$. After support-based pruning, we identify the candidates that has the highest support and look at the items they contain for the extraction of high confidence association rules. In fact, for the purpose of our application, when the support for each candidate is counted and tested against the maximum value possible instead of the *minsup* threshold, we either get the voting process, or a potential candidate itemset for the extraction of high confidence association rule.

Rule generation : The objective is to extract all the high-confidence rules from the frequent itemsets found in the previous step. These rules are called strong rules. - An association rule can be extracted by partitioning the itemset Y into two non-empty subsets X and Y-X such that X→Y-X satisfies the confidence threshold, *mincon*. - Each k-itemset, Y, can produce up to 2^k-2 association rules, ignoring rules that have empty antecedents or consequents (0→Y or Y→0). We determine the confidence of each rule by using the support counts as in Eq. (7). Confidence-based pruning compares rules generated from the same frequent itemset to consider rules that have higher confidence. Rules with low confidence are pruned.

In sum, with a large dataset of transactions, we can better understand the prediction models

if we look at the association rules. We implement the rules on board of the vehicles with the aim of getting a more efficient and reliable evaluation of the cause of congestion.

D. Back-Propagation algorithm

The methods we propose are compared to the Back-Propagation algorithm (BP) introduced in the literature [83], where vehicles transfer messages only if they have total knowledge about the cause. In this section, we will briefly describe the rules of BP for the sake of readability. The algorithm activates a process that shares the individual information collected by vehicle i in the following sequence :

- Vehicle i continuously broadcasts and receives BEACON messages from neighbours ;
- Vehicle i knows its current road segment and computes current travel time on the segment ;
- Update of traffic data on board of Vehicle i (Trajectory speed, travel time, Demand and gap between vehicles) ;
- If the observed travel time is above a threshold, Vehicle i creates a feature vector and predicts with a classifier the cause of congestion ;
- If there is no stored event in the database concerning this event, the vehicle creates and propagates backwards a message called Event Request (RQ). RQ is transmitted upstream via broadcast to all cars in its communication range, and allow to retain the event info locally on the segment for a minimum duration before propagating it to adjacent segments ;
- Otherwise if the duration is not reached, then the vehicle stores the RQ, because communication between vehicles on the same segment has to happen for a certain duration before propagation of the event to first order adjacent segments can be done ;
- Once the duration is reached, Vehicle i propagates backwards the Event Response (RP) message to adjacent road segments. RP is the message used to send the non-recurrent congestion event to adjacent segments after the duration expires.

The implementation of the VP, BF, DAT and BP methods in real conditions is presented in the following section.

5.4 IMPLEMENTATION AND RESULTSl

The proposed methods to elaborate a decision concerning the cause of urban congestion are tested and compared through different scenarios described next.

5.4.1 Simulation outline

The TAPAS Cologne scenario is assumed to be one of the largest traffic simulation dataset [45]. It covers the main road network within the inner city of Cologne. Demand mobility data traces for the 6-8am peak hours are provided. We create extended scenarios mounted on top of the base scenario to model atypical traffic conditions such as weather, incident, work-zone, special event and bottleneck. We create them using SUMO, a microscopic traffic simulator for the simulation of urban mobility [72]. To simulate an Incident/Work-zone, on the base scenario, we stop on a lane some vehicles for a specific amount of time. We vary the position on the edge and the duration. In inclement weather, which lead to decreases in the vehicles' velocities and a more careful and defensive driver behaviour, we change the parameters of the car-following model in the simulator. To simulate a special event, we generate trips to a particular destination, with random departures and random routes. We use a Poisson process to generate random timings for trips. The rate parameter λ is the demand per second from different sources. To generate random routes, given trips are assigned to respective fastest routes according to their departure times and a given travel time updating interval by SUMO's traffic assignment model. Table 6.1 contains a description of the experiments in each scenario used for synthetic training set generation.

In total, 24 experiments are investigated in this case study. Data of independent vehicles passing on the congested segments of each experiment in each scenario are evaluated. Transactions are created by the vehicles and later put into a supervised dataset for learning. The training dataset is a matrix with rows corresponding to transactions and columns to items. A data sample belongs to a target variable and each of the 6,970 data samples is thus represented by equal number of items. The comparative analysis is presented below.

5.4.2 Comparative analysis

The methods are validated by three indicators; estimation accuracy, detection time and percentage of false alarms. Accurate estimation of the root cause of congestion will enable road authorities to make more informed decisions about how to best reroute traffic. Also, lower detection time and false alarms will permit a rapid and exact reaction to resolve the traffic condition.

5.4.2.1 Voting Procedure

In SUMO, we simulated two scenarios of congestion due an incident and a weather condition with experiments 1.1 and 3.1. We generate urban mobility traces from the scenarios for usage

Table 5.1 Description of experiments

Scenario	Experiment#	Description
Incident	1.1	- Incident at the beginning of a lane
	1.2	- At the middle of a lane
	1.3	- At the end of a lane
	1.4	- For short duration
	1.5	- For long duration
	1.6	- Incident inside Impact Region
	1.7	- Outside Impact region
Workzone	2.1 - 2.8	- Similar to incident experiments
Weather	3.1	- Heavy weather condition
Special Event	4.1 - 4.4	- Four different ingress flows
Bottleneck	5.1 - 5.4	- Four junctions on the base scenario

in ns-2, the discreet-event network simulator. In the simulation of vehicular communications, we assume that vehicles are equipped with a Global Positioning System (GPS) device for positioning, a transceiver for communication using Dedicated Short-Range Communications (DSRC), and an enriched digital road map containing information about the map. We use well known data forwarding techniques to pass information through the CVs such as geographical routing (Geocast), and broadcast. For communication among all cars, we assume standard signal range of the 802.11p protocol, which is 300 meters. BEACON messages are exchanged every 0.1 seconds.

Fig.5.5 compares the average percentage of vehicles accurately estimating the cause of congestion in each scenario between VP and BP method. The best results the BP can do are after a certain time has elapsed because according to the algorithm, the first minutes following an excessive congestion not all vehicles on the segment are going to exchange messages regarding the event. On the other hand, at 4200s, vehicles in an incident start to vote via the VP. Only a few would have assessed that excessive congestion is present and will vote about the cause. Votes of all other vehicles on the segment are not counted because they did not have a probabilities vector for the estimation of the cause. At 5700s, when almost 62% of the vehicles in VP voted about the cause of congestion and all vehicles on the road segment were aware of the cause of congestion for more than 1500s, the BP algorithm activates and start signalling to vehicles that congestion is due to an incident. After 5700s, the accuracy in VP is almost the same as that in BP in this scenario. We also notice that the increasing speed of

the percentage of vehicles in VP is smaller than that in BP because according to the BP algorithm, after the duration has elapsed, vehicles experiencing excessive congestion propagate the message backwards if they are in the communication range of each other and regardless if all others finally experience excessive congestion or not. Similarly, in the weather scenario at 5220s vehicles vote via the VP and it's only at 5820s, when almost 52% of the vehicles in VP estimated the cause of congestion, that the BP algorithm activates and start signalling to vehicles that congestion is due to weather. We simulated the scenario of an incident and a weather condition on 15 other road segments and found that the VP outperformed the BP method in every experiment in terms of percentage of vehicles accurately estimating the cause of congestion. On average, it did so by approximately 48%. This increase in performance is due to the fact that the VP method use CVs to elaborate a decision collectively on the cause of congestion. This shows that cooperation on a given geographical area to obtain deeper insight of traffic condition improves the accuracy of the estimation in contrast to BP that used CVs to disseminate the information and not for the sake of evaluation.

To study the performance of the methods in terms of detection time, firstly we show in Fig.5.6 how the parameters of traffic flow change with time in the urban road network on a particular edge. The flow is zero either because there are no vehicles and density is zero or there are too many vehicles so that they cannot move and it is maximum density. Vehicles observe their travel time on the segment in order to detect congestion. It's only when vehicles detect congestion that they analyze if it is excessive. In case the observed travel time on a road segment is excessive, vehicles assume they are experiencing non-recurrent congestion and estimate the cause. In Fig.5.5, VP outperforms BP in terms of detection time because the algorithm of the BP requires that vehicles exchange their evaluation only if they experience the excessive congestion for a certain duration of time and they are in the communication range of each other. In the VP, vehicles on the road segment are involved in the exchange as soon as one vehicle detects the excessive congestion and triggers the procedure, even if the other vehicles don't yet experience it. Precisely, we observe from the figure that detection time is reduced from 5640s to 360s for the incident scenario and for the weather condition, reduction corresponds to 10.3%.

On the other hand, in Fig. 5.7 we compare the methods with the percentage of false alarms in five different scenarios. A false alarm is a vehicle initiating a VP or a BP method and the simulation shows no excessive congestion, i.e. the situation captured by each vehicle is compared with the real simulation. We perform two experiments on each scenario. In the first experiment vehicles execute the BP procedure and in the second the VP. On 200 different road segments, we monitored vehicles passing and we report on average the percentage of false alarms triggered by the BP and VP. We found an average percentage error ranging

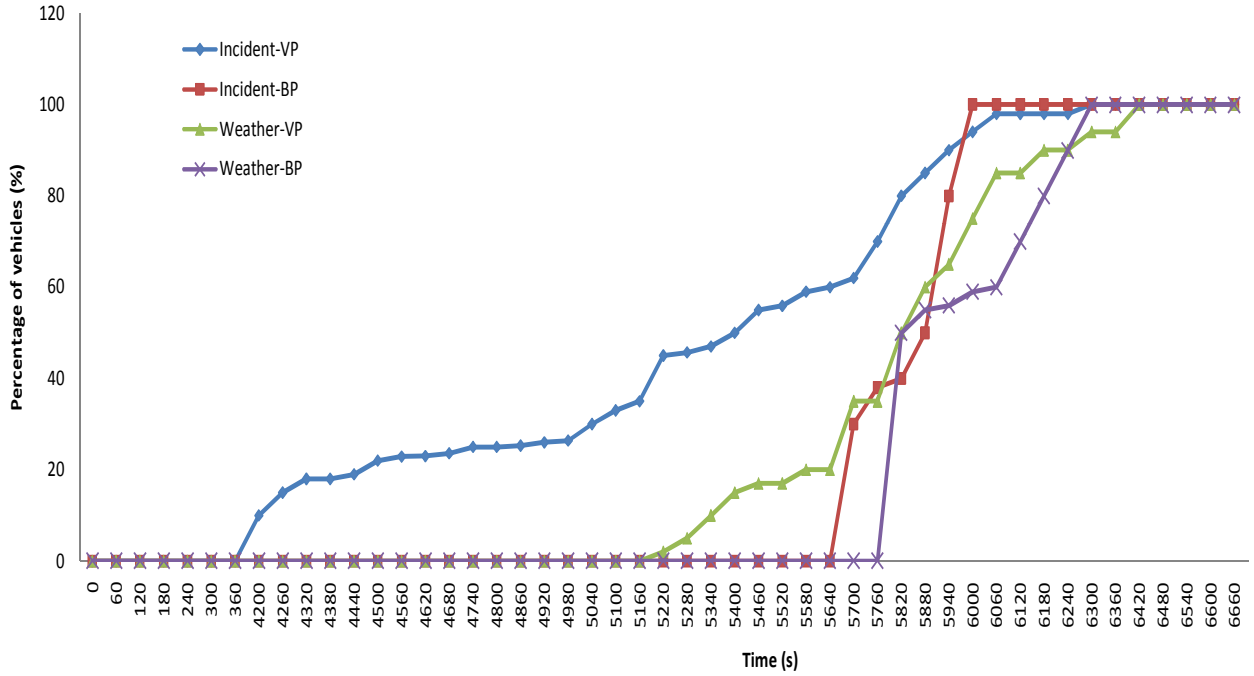

Figure 5.5 Voting Procedure - Percentage of vehicles accurately estimating the cause of congestion in different scenarios

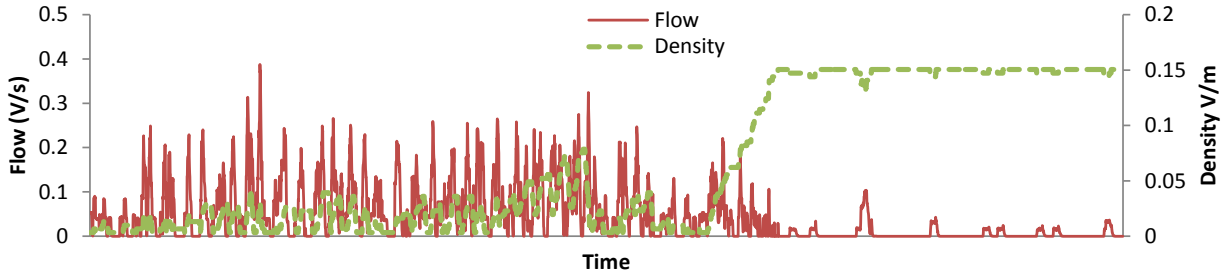

Figure 5.6 Variation of the parameters of traffic flow on an edge

from 3% to 11%. Although VP outperforms BP in terms of detection time and estimation accuracy, we notice in Fig. 5.7 that the voting procedure has the highest percentage of false alarms.

To investigate the situation further, we notice that VP detects the cause of congestion earlier than BP in every situation where the real simulation represented excessive congestion. The

problem is that VP triggers false alarms in the other scenarios, where there was no excessive congestion. This is due to the fact that VP lacks a procedure that is able to control situations where the assessment of a few vehicles do not represent a statistically compliable state. In fact, voting is triggered as soon as one vehicle detects excessive congestion. BP handles false alarms in a better manner than the VP because the BP algorithm requires that vehicles in the communication range of each other be aware of the same situation for a certain duration of time before triggering the back propagation of the events and declaring that the road segment is congested. We try to solve this problem with the theory of belief functions. The theory adds knowledge to the messages exchanged in order to decrease false alarms and we present it in the next section.

5.4.2.2 Belief functions

The VP is an improvement of the BP presented in terms of estimation accuracy and detection time. Nonetheless, if vehicles vote, some vehicles cannot quantify their ignorance on the presence or the absence of congestion. In other words, their vote is not an accurate representation of information. Vehicles cannot quantify their ignorance on the presence or the absence of congestion. With belief functions partial knowledge or total ignorance can be represented.

We simulated a scenario of a congestion caused by inclement weather. Vehicles exchange at the same time period messages assigning a mass for the singleton, another for a subset and an ignorance degree correspondingly. We report in Table 5.2 results of vehicles S1-S22 exchanging messages M1-M22 at time t=5580s. Their mass functions have been combined using the conjunctive rule of combination resulting with a high confidence degree of 0.85 in the congestion caused by a weather event, and a low ignorance degree, with no conflict.

For the same scenario, we collect messages exchanged by the BF method during 1980 seconds after congestion is detected. Fig. 5.8 shows the estimation accuracy of the BF compared to the VP and BP.

Results show that the BF method gives the best estimation of the cause of congestion. Specifically, BF outperforms the BP method by approximately 58%. We also notice the time elapsed before vehicles on a segment decide about the cause of congestion. The classification algorithm infers on the cause of congestion at 5220 seconds, only when there is congestion on the road segment and when this congestion becomes excessive. In terms of detection time, the VP and BF methods outperform the BP by 9.4%. In fact, the BF and VP methods detect the cause of congestion 560 seconds before BP. The reaction time of the BP is slower because the algorithm requires that vehicles transfer in the vicinity their assessment only when they have total knowledge about the cause of congestion and that a certain duration has elapsed.

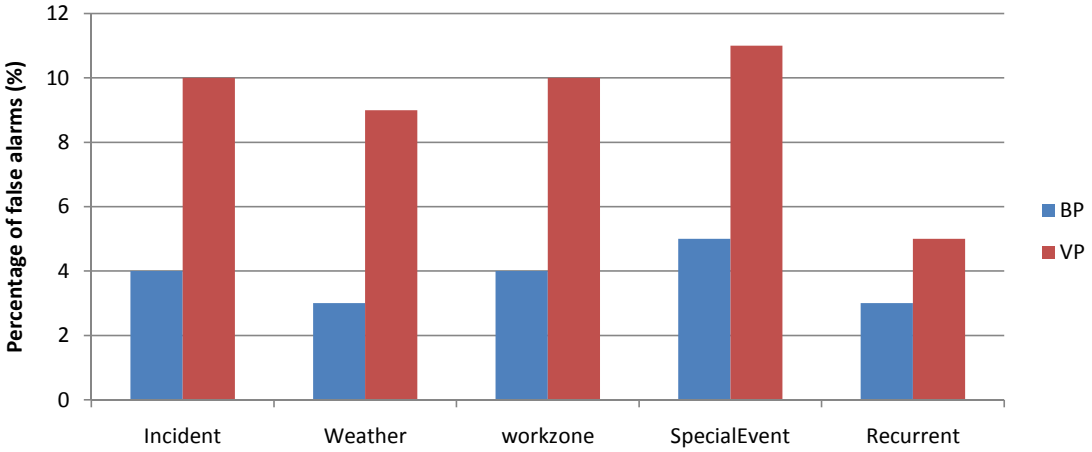

Figure 5.7 Percentage of false alarms of the VP and BP

Table 5.2 Combination of mass functions m_i from messages M_i,i ∈ {1,2,3,...,22}

Hypothesis	m_1	m_2	m_3	m_4	...	m_{21}	m_{22}	$m_{1\cap2\ \cap\ ...\ \cap22}$	BetP
∅	0	0	0	0		0	0	0.652	
Incident (I)	0	0	0	0		0	0	0.022	0.04
Weather (We)	0.4	0.62	0.7	0.6		0	0.67	0.1637	0.85
Workzone (Wo)	0	0	0	0		0	0	0	
Special Event	0	0	0	0		0	0	0	
Recurrent (Re)	0	0	0	0		0.61	0	0.1068	0.11
{I or We}	0	0	0.2	0.1		0	0.3	0.0234	
{I or Wo}	0	0	0	0		0	0	0	
{I or SE}	0	0	0	0		0	0	0	
{I or Re}	0	0	0	0		0	0	0	
{We or Wo}	0	0	0	0		0	0	0	
{We or SE}	0	0	0	0		0	0	0	
{We or Re}	0.3	0.3	0	0		0.34	0	0.032	
{Wo or SE}	0	0	0	0		0	0	0	
{Wo or Re}	0	0	0	0		0	0	0	
{SE or Re}	0	0	0	0		0	0	0	
Ω	0.3	0.08	0.1	0.3		0.05	0.03	0.0000011	

In the VP and BF methods, vehicles share their partial knowledge and the voting or the belief functions conclude earlier about the cause having the highest probability.

In Fig. 5.9 we monitor the methods with the percentage of false alarms in five different scenarios. Compared to the VP method, BF decreases the percentage of false alarms by approximately 1.8%. In the recurrent scenario, BF yields the same performance as the BP method. This shows that in the evaluation process, not only cooperation between vehicles but adding knowledge to the messages exchanged improves the performance. Nonetheless, BP still outperforms BF by approximately 4.25% less false alarms in the incident, weather, work-

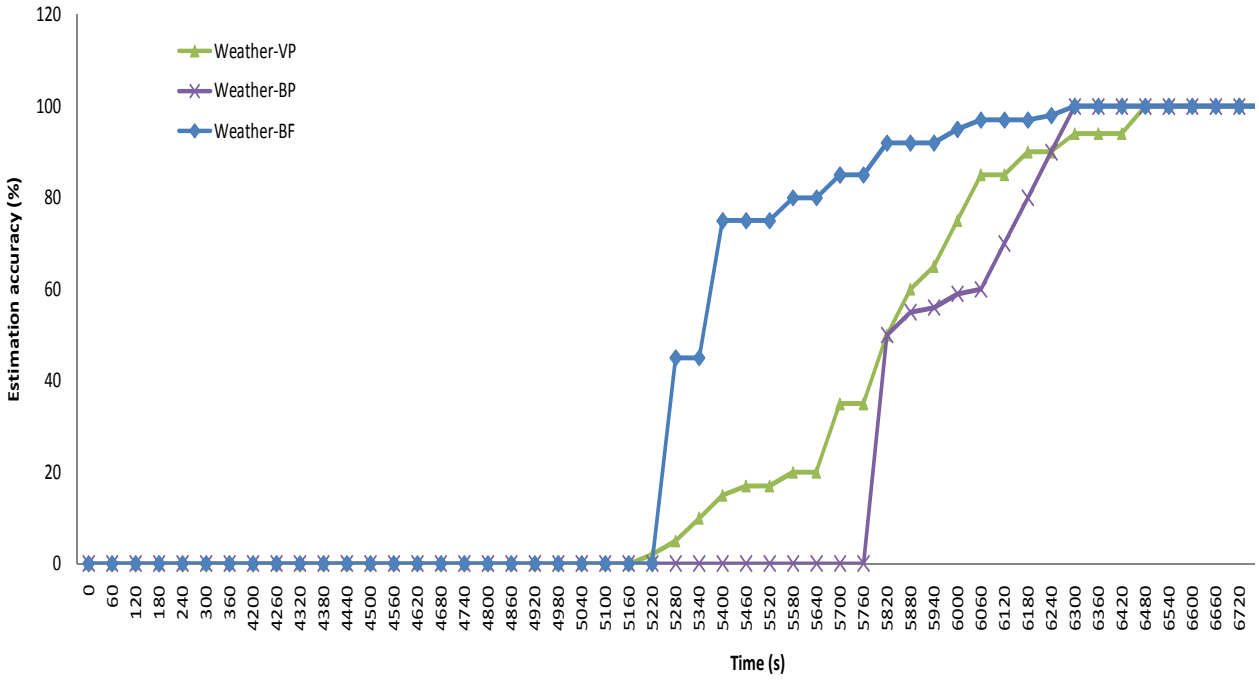

Figure 5.8 Estimation accuracy of different methods in a scenario of congestion caused by weather

zone and special-event scenarios. To add more knowledge on board of each vehicle, in the next section we present the implementation and results of the data association technique. We explore data the vehicles collected for learning purposes by building a dataset and extracting relationships via a data mining technique to extract more knowledge.

5.4.2.3 Data mining technique

We analyse the messages exchanged by CVs for frequent patterns in order to identify relationships for rule generation. We use the default settings of the Apriori principle implemented in Weka [79], a data mining software, to generate the rules. We then proceeded to the evaluation of the association rules. Association rule algorithms tend to produce a large set of rules, many of them are 'uninteresting'. To determine the interestingness, we consider the subjective interestingness measure. A rule is considered subjectively uninteresting unless it reveals unexpected information about the data, or provides useful knowledge that can lead to profitable actions. This measure is defined based on domain information. As domain experts, we extract the rules of interest because support and confidence measures are insufficient at filtering out uninteresting association rules. We present in Table 5.3 the rules extracted from

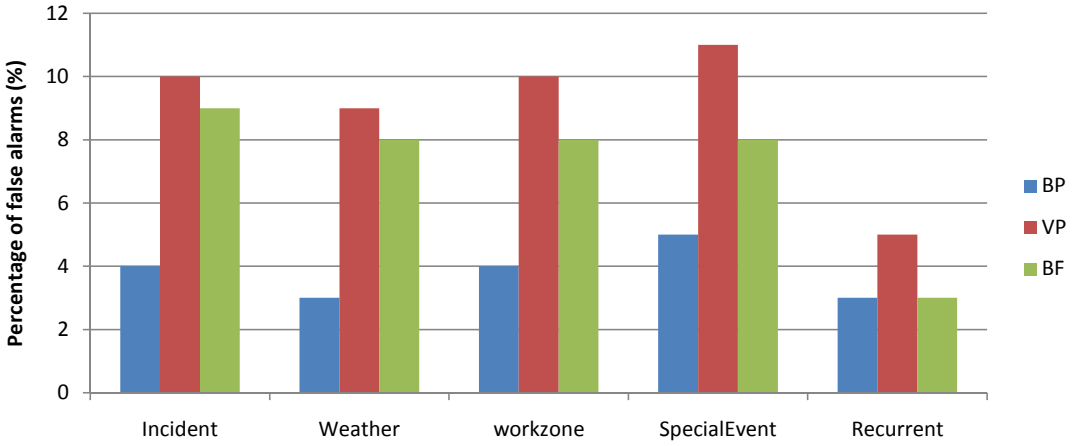

Figure 5.9 Percentage of false alarms of the BF, VP and BP methods

the supervised dataset.

Rules 1 to 9 are learned from the dataset. Rules 1 and 2 are extracted from the mining of the ordered items in each transaction, and we seize that when vehicles are experiencing congestion caused by a special event, their second guess will never be an incident or a work-zone. We considered only the first and second item in each transaction because of the subjective interestingness measure. Similarly, for rules 3 to 6, we understand the guess of vehicles in any scenario. Differently, rules 7-8-9 are extracted from the first item in each transaction and the label of the transaction. The consequent part of the rule is the label. Because each transaction is created by a vehicle in a specific scenario, the label of the transaction is the scenario. If the label is different from the first item in the transaction, it means that the vehicle wrongfully predicted the cause of congestion. By mining the dataset we uncover those rules that when applied in the DAT method, enhance the accuracy of prediction of the cause of congestion. We consider the scenario of congestion occurring on a road segment due to recurrent traffic. Fig. 5.10 shows the percentage of vehicles accurately estimating the cause of congestion with different methods.

In the VP and BF methods, vehicles decide cooperatively without applying the association rules. When applying the mining rules in DAT, performance is greater in terms of estimation accuracy. BF model's partial knowledge allowing earlier detection of the cause and DAT gives further precision on incoherencies in the data having the best estimation. In the DAT experiment, vehicles make use of the belief functions and association rules to estimate the cause of congestion. The DAT improves estimation accuracy of 71% compared to the BP method. In this scenario, detection time is decreased by 9.45% informing of the congestion

Table 5.3 General association rules

Rules	Frequency	Description
1. SE→ We	50%	If a vehicle predicts that the main cause of congestion is a SE, then 50% of the time the second guess is that the cause might be We.
2. SE→ Re	50%	If a vehicle predicts that the main cause of congestion is a SE, then 50% of time the second guess is that the cause might be Re.
3. We→ Re	60%	If a vehicle predicts that the main cause of congestion is a We, then 60% of the time the second guess is that the cause might be a Re.
4. We→ I	40%	If a vehicle predicts that the main cause of congestion is a We, then 40% of the time the second guess is an incident.
5. I→ SE	100%	If a vehicle predicts that the main cause of congestion is an Incident, then 100% of the time the second guess is that the cause might be a SE.
6. Wo→ SE	100%	If a vehicle predicts that the main cause of congestion is a Wo, then 100% of the time the second guess is that the cause might be a SE.
7. We,SE-→ We	100%	If some vehicles on the road segment predict that the cause of congestion is due to We and others on the same segment predicts it is due to a SE, then the cause of congestion is always due to Weather.
8. Re,SE-→ Re	100%	If some vehicles on the road segment predict that the cause of congestion is due to recurrent traffic and others on the same segment predicts it is due to a SE, then the cause of congestion is always recurrent.
9. Re,We-→ Re	100%	If some vehicles on the road segment predict that the cause of congestion is due to weather condition and others on the same segment predicts it is recurrent, then the cause of congestion is always recurrent.

cause earlier. Created messages with the BP method have a confidence equal to 100%, and they are transferred depending on the evaluation of the cause of congestion done in each vehicle in the range of communication of the vehicle that initiated the backpropagation. For these reasons, this method gives poor results, but later in the simulation, BP tends to better results and that is only when there are enough vehicles adequately positioned in the communication range of the initiating vehicle and having the same evaluation as the initiator.

In Fig. 5.11, we compare the methods with the percentage of false alarms in five different scenarios. Compared to the BF method, DAT decreases the percentage of false alarms by approximately 1.33%. This shows that the association rules on board of each vehicle add knowledge and improve the performance. Nonetheless, BP still outperforms DAT by approximately 3.25% less false alarms in the incident, weather, work-zone and specialevent scenarios.

To investigate the situation further, we specifically present in Fig. 5.12 the detailed monitoring

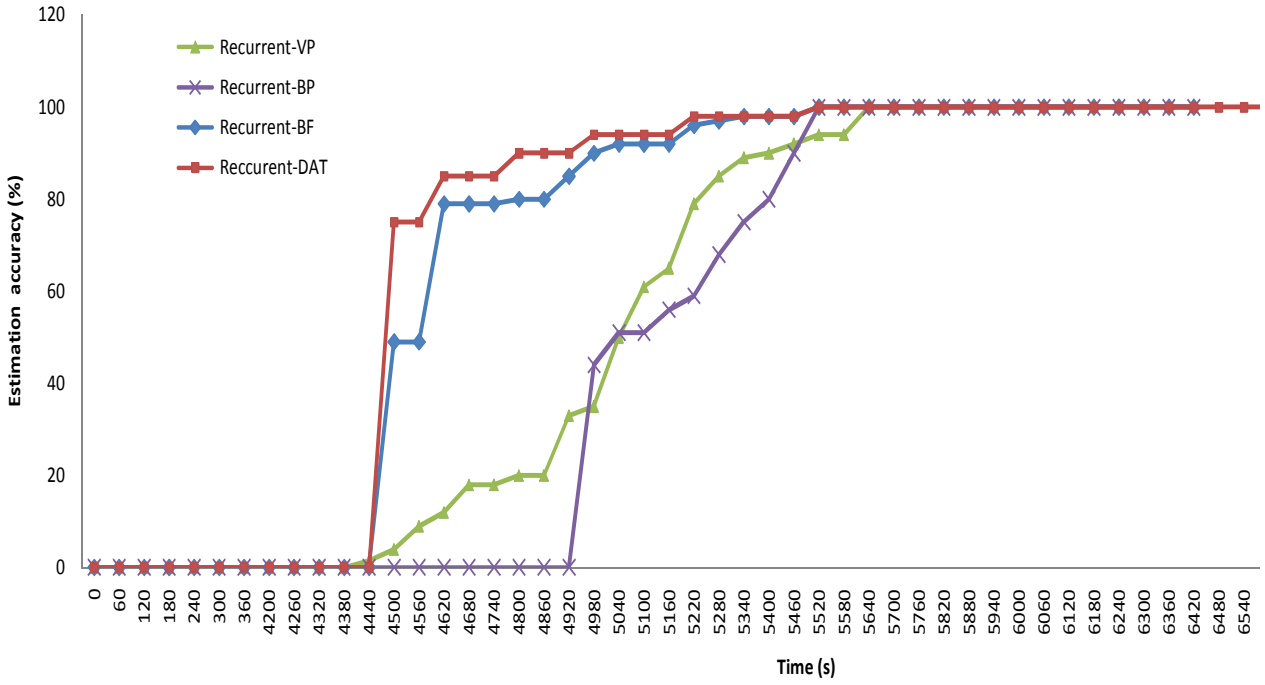

Figure 5.10 Comparative estimation accuracy of vehicles when congestion is due to recurrent traffic

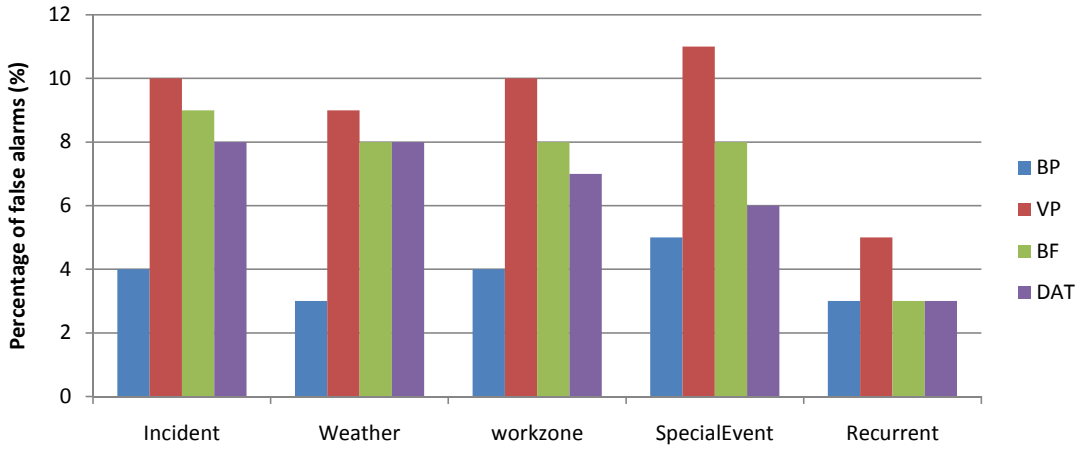

Figure 5.11 Percentage of false alarms of the BF, VP, BP and DAT methods

of the incident scenario for the VP and BP experiment. We observe that in both VP and BP, false alarms happen earlier in the simulation and dissipate as congestion installs. In fact, in a dense environment, where vehicles inform rapidly about events, the percentage of false alarms is higher at the early stage of the experienced congestion. The percentage decreases

after a certain period and tends to zero as simulation time advances.

We make use of this information and adapt the DAT method with an addition of a time factor, β. We call this method the β-DAT where we force the vehicles to wait for a certain time period, β, before cooperating for the evaluation of the cause of congestion. We found that the value of β has a direct impact on the percentage of false alarms and detection time and that it has no impact on the estimation accuracy. If the value of β is high, vehicles take as much time as BP to detect congestion and the percentage of false alarms tends to zero. With an upper bound for β being the duration of four consecutive LJT (Link Journey Time) as in [10], we conducted experiments and found that half of that duration is enough to attain the expected performance.

We present in Fig. 5.13 the performance of the β-DAT method in terms of estimation accuracy and detection time for the scenario of congestion occurring on a road segment due to recurrent traffic. We notice that, similar to the DAT method, β-DAT improved estimation accuracy by approximately 70% compared to the BP method. Also, detection time of β-DAT is 7.09% lower than that of the BP method, informing of the congestion cause earlier. It's 2.36% higher than the DAT method, a slight increase of approximately 120 seconds. The consequence of adding a time factor to the DAT method on the detection time is insignificant compared to the benefit the duration added to the percentage of false alarms. In Fig. 5.14 we compare the methods with the percentage of false alarms in five different scenarios. We see that β-DAT has the lowest percentage of false alarms in all scenarios. In fact, compared to the DAT method, β-DAT decreases the percentage of false alarms by approximately 3.6%. Also, β-DAT outperforms BP by approximately 1.25% less false alarms triggered by the network on the road segment. This shows that adapting the duration in combination with cooperation between CVs and knowledge on board of each vehicle improves overall performance for the accurate estimation of the cause of congestion.

5.4.3 Penetration rate of CVs

In this section, we present the results of the market penetration rate on the performance of the methods. Specifically, we show the impact of the percentage of vehicles equipped with VANET technology on the estimation accuracy and detection time. We vary the number of participating vehicles in the incident scenario and calculate the percentage of vehicles accurately estimating the cause of congestion as a function of time. We report the results in Fig. 5.15.

A low penetration rate implies that few vehicles are equipped with transceivers and this leads to network fragmentation in the context of ad hoc networks. As a consequence, in

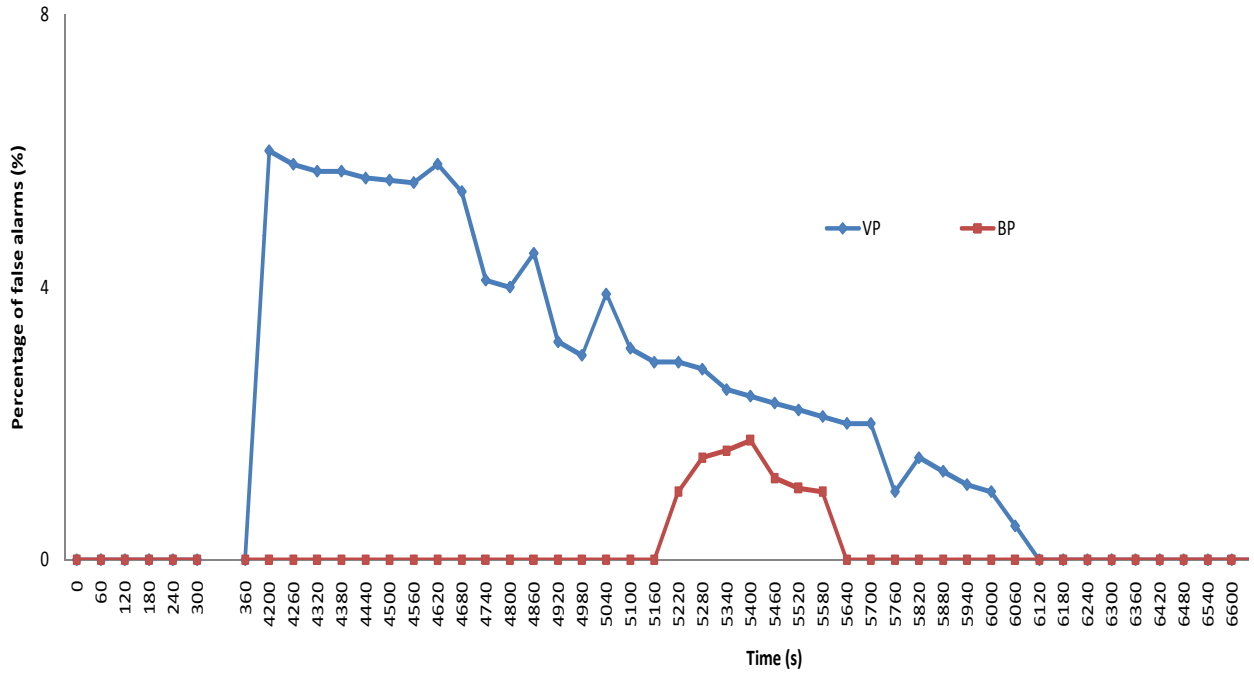

Figure 5.12 Monitoring of false alarms in the incident scenario

the BP method vehicles have to wait to be in the communication range of each other for the algorithm to conclude on the cause of congestion. As there is no other exchange of traffic information, its performance strongly depends on the penetration rate of participating vehicles. In the figure, the estimation accuracy of BP is very low and detection time is very high for penetration rates of 10%, 50% and 75%. On the other hand, in the VP, BF, DAT and β-DAT, they need not to be in the communication range of each other to conclude on the cause of congestion but rather to collect the evaluation done by each other. Fragmentations in these methods have a lower impact on the performance because the methods do not depend on network connectivity to resolve. This is shown in the figure by the percentage of vehicles being very close to that of a penetration rate of 100%. We only present the performance of the VP because the other methods present the same tendency. We notice that for a penetration rate of 10% and 50% in the VP method at 5460s and 5340s respectively, the percentage of vehicles accurately estimating the cause of congestion is even slightly better that of a penetration rate of 100%. This is due to the fact that in this particular scenario, if few vehicles are equipped with transceivers and they vote accurately about the cause of congestion, the ratio is at times better than that of more vehicles equipped with transceivers voting inaccurately about the cause of congestion. Overall, in other scenarios of incidents,

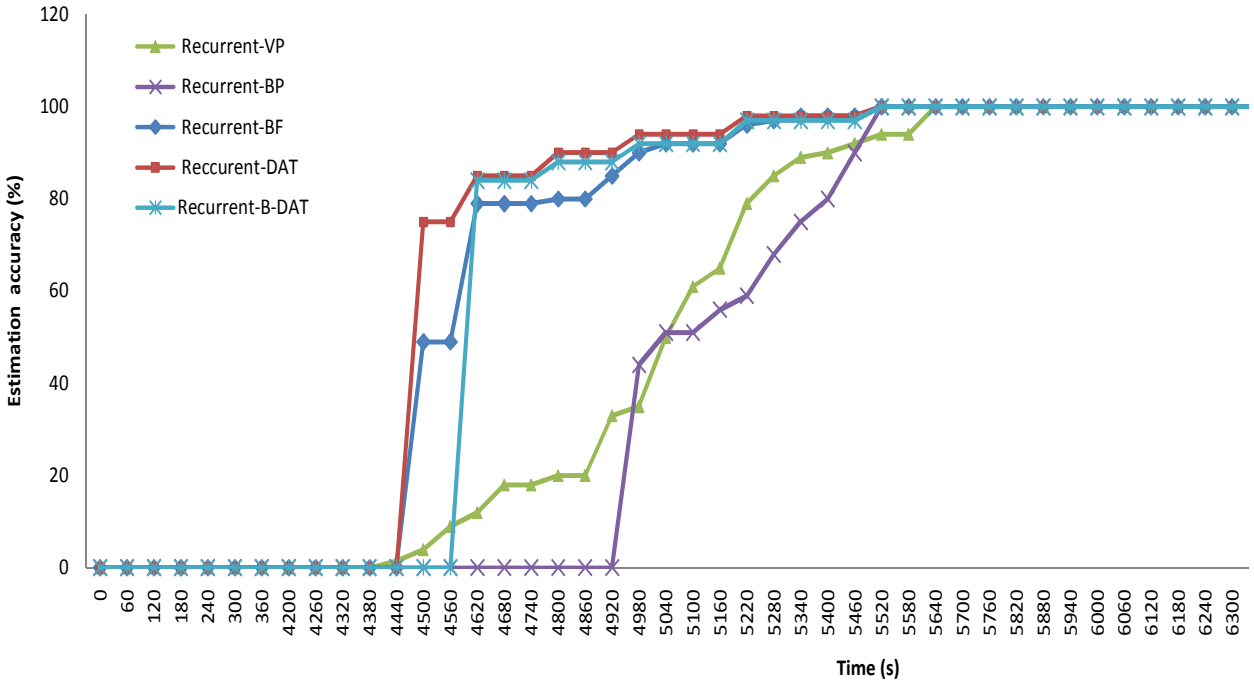

Figure 5.13 Performance of β-DAT for a scenario of congestion due to recurrent traffic

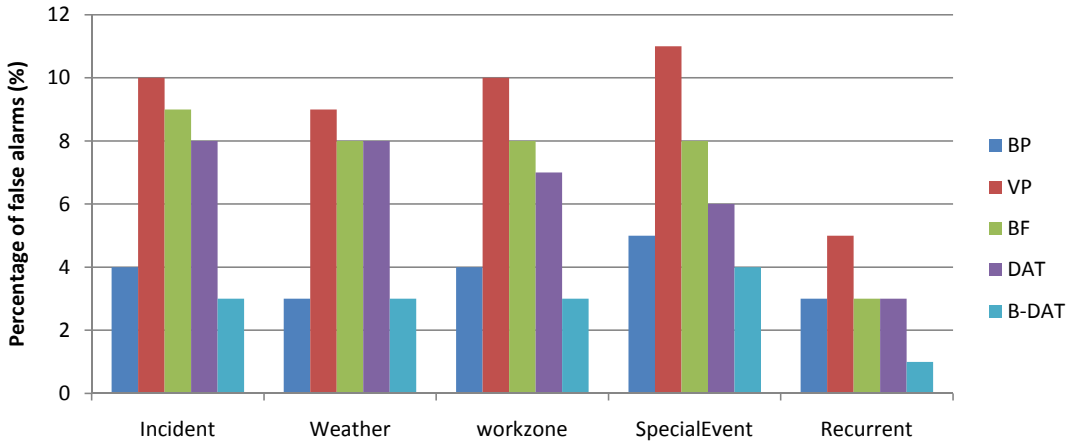

Figure 5.14 Percentage of false alarms in different methods

the trend is the same and estimation accuracy follow the curve of 100% penetration rate. However, we see from the figure that a low penetration rate affects detection time more than estimation accuracy. In fact, at 10% and 50% penetration rate, vehicles detect that congestion is due to an incident only at 5460s and 5040s respectively. The question that arises in such a

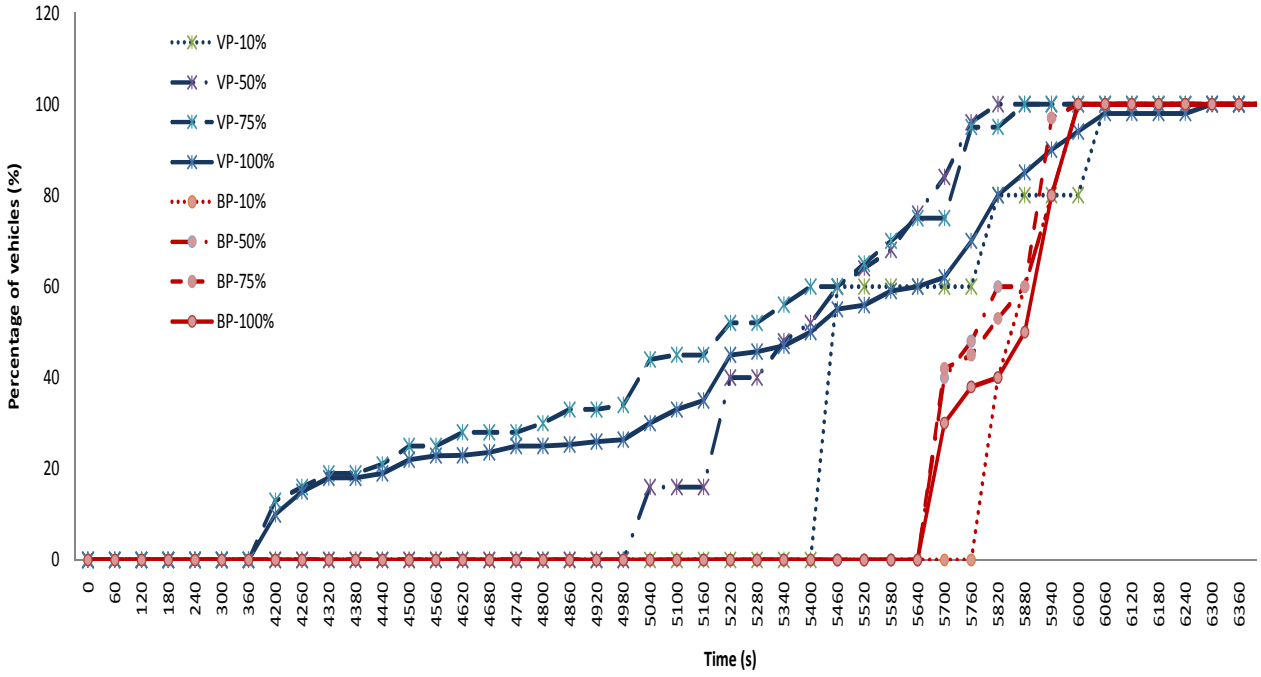

Figure 5.15 Impact of penetration rate on the performance of the methods in the incident scenario

case is what percentage of penetration is enough to obtain the full benefits of the methods. To study the necessary market penetration of V2V communications technologies to reach in order to get the best performance in terms of estimation accuracy and detection time, we conclude from the figure that at 75% of penetration rate, detection time and estimation accuracy show the best performance and constitute an upper bound. The methods showed the same performance in other scenarios of congestion due to a work-zone, special event and recurrent congestion. Penetration rates of 75% showed the best performance in all scenarios. On average, approximately 63% of penetration rate was acceptable in all scenarios except when the cause of congestion is a weather condition.

In case of congestion caused by inclement weather, a very high percentage of CVs is required in order to achieve good performance. To study the situation further, we note the effect of weather on mobility. Firstly, Fig. 5.16 illustrates *g(t)*, the gap, the empty space after the leading vehicle along a vehicle's trajectory in the urban network in normal conditions. The following distance mainly depends on the speed of the following vehicle which is adapted to the speed of the leading vehicle. The desired following distance between two consecutive vehicles is highlighted in Fig. 5.16 and corresponds to the minimum safe gap attained by

the vehicle. Our analysis starts with the collection of desired gaps in a scenario of inclement weather, creation of cumulative gap distribution and calculation of the *85th* percentile gap observed. In Fig. 5.17, we show the *85th* percentile gap measures collected by vehicles in the normal scenario and the weather scenario. Incident, work-zone, special event and bottleneck scenarios showed values that are similar to the normal scenarios, thus, they are not presented. From the figure, we see that in inclement weather, most vehicles augment their following distance in comparison to normal weather conditions.

We conclude that the lower performance is not only due to the fact that very few vehicles are equipped with transceivers, but also that in inclement weather the gap between vehicles is increased, and this leads to more network fragmentation than in other scenarios.

5.5 Conclusion

Given that traffic involves multifaceted complex interactions, exploring the cause of congestion at a vehicle level is a partial limited solution because currently, each vehicle classifies individually the cause of congestion based on its personal trajectory. The assessment and classification are done locally on a vehicle level and in the event of a false alarm, in terms of security, spreading uncertainty among vehicles or false information causes more congestion, disrupts the proper network operation and presents a serious challenge. We proposed methods to obtain deeper insight on the cause of traffic congestion using cooperation between CVs since information is a subject of interest to vehicles in a given geographical area. Besides cooperation, we proposed that an evaluation process has to take place after data sensing and before data fusion and aggregation. We added this layer to address the vulnerability of fusion algorithms and also to lower the side effects of false alarms because their impact is comparable to security threats to the network. Finally, we explored the collected data for learning purposes by building a dataset and extracting relationships and knowledge via data mining techniques to elaborate a decision collectively. We also studied the impact of the market penetration rate of CVs on the performance of our methods.

We have considered a realistic map configuration of the city of Cologne in the evaluation of our methods. We tested and compared the methods using a microscopic urban mobility simulator, SUMO and a network simulator, ns-2, for the simulation of communication between CVs. Compared to the backpropagation (BP) technique proposed in the literature, we have obtained an enhanced estimation accuracy of 48% for the voting procedure (VP), 58% for the belief functions (BF), 71% for the data association technique (DAT) and 70% for the adapted data association technique (β-DAT). The methods also enhance the detection time by 10.3% for VP, 9.40% for BF, 9.45 for DAT and 7.09% for β-DAT. We monitored the

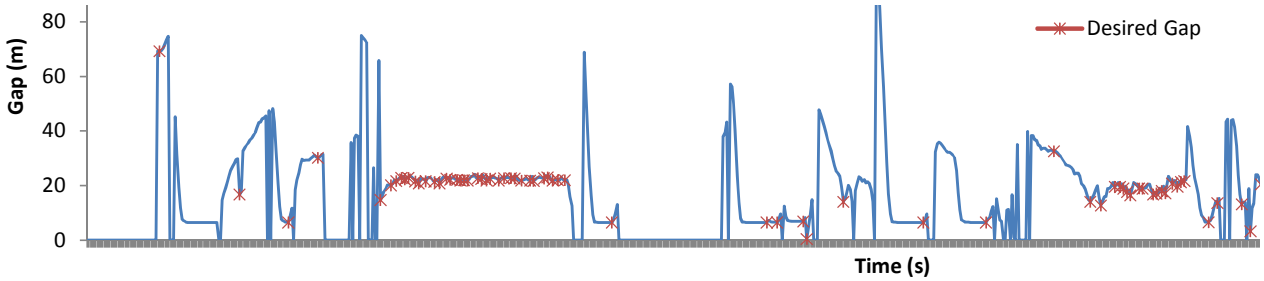

Figure 5.16 Following distance of a moving vehicle in the base scenario

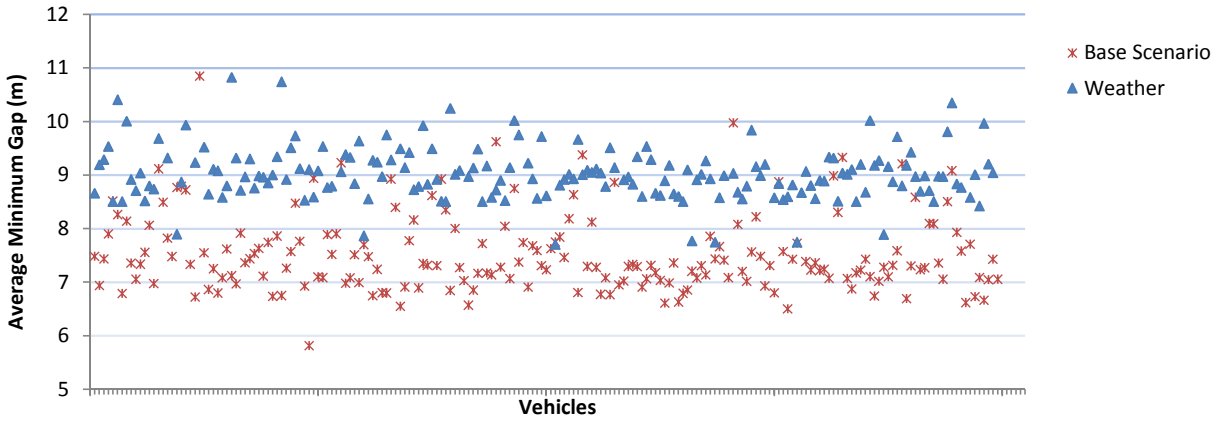

Figure 5.17 85th percentile gap values of vehicles in different scenarios

methods with the percentage of false alarms and β-DAT outperforms BP by approximately 1.25% less false alarms triggered by the network. This shows that adapting the duration in combination with cooperation between CVs and knowledge on board of each vehicle improves overall performance for the accurate estimation of the cause of congestion.

Finally, our work has shown that our methods require approximately 63% penetration rate to ensure satisfactory performance and obtain the full benefits of V2V communications technology for accurate estimation of the cause of congestion. We understand that traffic efficiency applications do not require such a high market penetration level of CV technology as is the case for safety-related applications. Based on the simulation results, automatic traffic controlling systems may use the methods to prevent traffic jams and introduce countermeasures as soon as the cause of congestion is detected. In the future, since scalability will profoundly impact the system performances, scalable schemes based on hierarchy or cluster should be

creatively elaborated because the proposed methods for congestion detection and classification are not scalable in a large scale. Also, robustness has to be considered because it makes vehicles aware of the overall congestion level on a particular road segment, despite the short-term changes in vehicle's mobility. Finally, these systems cannot be deployed in the near future, as one has to wait until the necessary market penetration of V2V communications technologies has been reached. Since our methods are based on an infrastructure-less approach, it could be interesting to investigate how much of a lower penetration rate can be achieved if the same methods were adapted to an infrastructure-based approach. The drawback of infrastructure-based is that service charges will most probably apply. However, they can be deployed in the near future.

The evaluation phase, cooperation process and rules generated by our scheme make our approach generalizable and portable to other cities and networks. However, it could be interesting to validate our scheme with data of connected vehicles in real-world conditions. Unlike the synthetic dataset in our work, it will provide real world training dataset in occurrence of different non-recurrent congestion scenarios and then the results of the data mining technique can be compared to our work.

CHAPTER 6 ARTICLE 3 : PREDICTION OF TRAFFIC FLOW VIA CONNECTED VEHICLES

Ranwa Al Mallah, Alejandro Quintero, and Bilal Farooq
submitted to IEEE Transactions on Intelligent Transportation Systems

Abstract

Urban traffic congestion is growing at an alarming rate. We propose a Short-term Traffic flow Prediction (STP) framework so that traffic managers take early actions to control the flow and prevent the congestion state. We anticipate flow at future time frames on a target road segment based on historical flow data and innovative features such as real time feeds and trajectory data provided by Connected Vehicles (CVs) technology. To cope with the fact that existing approaches do not adapt to varying traffic situations, we show how this novel approach in this domain allows advanced modelling by integrating the impact of the various events that CVs realistically encountered on segments along their trajectory into the forecasting of flow. We solve the STP problem with a Deep Neural Networks (DNN), and tackle the problem by learning the target DNN in a MultiTask Learning technique (MTL). The results show our approach outperforms state-of-the-art ARIMA time series and baseline classifiers, with an average Root-Mean-Square Error (RMSE) of 0.05. Compared to single task learning with Artificial Neural Network (ANN), ANN had a lower performance, 0.113 for RMSE, than MTL. Because a transportation system is a highly correlated network, with characteristics such as large amounts of data and high dimensions of features, DNN learned historical similarities between road segments, in contrast to using direct historical trends in the measure itself, since sometimes trends may not exist in the measure but do in the similarities.

6.1 Introduction

Road traffic congestion is a particular state of mobility where travel times increase and more and more time is spent in vehicles. Apart from being a quite stressful experience for drivers, congestion also has a negative impact on the environment and the economy. In this context, there is pressure on the authorities to take decisive actions to improve the network traffic flow. By improving network flow, congestion is reduced and the total travel time of vehicles is decreased. To this end, predictive techniques are needed by infrastructure operators to

allow advanced modelling. The fast prediction of traffic flow on a road segment allows the traffic managers to take early actions to control the traffic load and prevent the congestion state [11]. Particularly, short-term prediction, STP, enables road authorities to make more informed decisions about how to best reroute traffic, change lane priorities and modify traffic light sequences. It may also assist in better planning of road network expansion, as well as optimal road sign placement and speed limit setting.

In the context of short-term traffic flow prediction, many studies have been devoted to highways rather than highly congested urban regions [47], [87], [88]. In a highway scenario, the road section can be modeled as a network flow model that require flow conservation on all segment. The amount of flow entering an arc equals the amount of flow leaving the arc. In an urban scenario, on the other hand, each arc of the underlying graph has an associated positive gain or loss factor. Flow passing through the arc is magnified or diminished by a factor. In fact, the problems faced in urban traffic are not easy to solve because they depend on multiple dynamic aspects that are difficult to describe and to model in detail. They are intricate, complex networks and far more likely to be monitored by the traffic authorities. Therefore, the design of accurate and scalable traffic flow prediction for urban road networks is required.

On another hand, several unresolved problems exist for the short-term traffic flow prediction on urban networks. A traffic management system must firstly ensure efficient monitoring of the urban network. Currently, traffic state cannot be directly measured everywhere on the traffic road network because infrastructure operators are strained to monitor traffic while using the least possible resources [89]. Current collection methods rely on dedicated traditional heterogeneous sensor and backbone networks and hardware/software solutions [90]. Operators interpolate information from incomplete, noisy and local traffic data because deploying highly sophisticated equipment to ensure the accurate estimation of traffic flows and timely detection of events everywhere on the road network is not the ideal solution due to the limitation in financial resources to support dense deployment and the maintenance of such equipment, in addition to their lack of flexibility. Due to the high complexity and uncertainty of contemporary transportation systems, these methods fail to capture in detail and in real time all the dynamics, they are not capable of evolving over time and do not scale to larger cities. Therefore, alternative cost-effective and flexible solutions are needed to guarantee better monitoring of road traffic.

A typical measurement that interest system developers for prediction is the flow rate on a target road segment, which is the number of vehicles that pass through a segment per time period. However, in addition to traditional traffic sensors that collect flow values on target

segments of the road network, a variety of data sources, such as lidar, radar, and video from surveillance cameras have emerged in traffic flow prediction research [91]. In fact, to increase prediction accuracy, scientists now recognize that the problem is that traffic flow prediction heavily depends not only on historical flow values on a segment but also on real-time traffic data. Those real-time feeds are presently collected from various sensor sources, including inductive loops, radars, cameras, mobile Global Positioning System, crowd sourcing, social media, and incorporated in the prediction of flow [13]. As the data originate from different sources, their conversion is the most important step. In this process, the first obstacle is the amount of data collected which is increasing exponentially, and the second is its complexity. This makes data conversion difficult and highly time and resource consuming. The next steps are relevant data extraction and cleaning, as well as data reduction. Each of these tasks has its own challenges, including defining what is relevant and what is noise, identifying one or the other, and extracting the useful data, given certain accuracy expectations. In sum, data aggregation poses many challenges when a variety of data sources are required in the process of data collection. advanced monitoring techniques should be deployed and must be capable of aggregating traffic data feeds from various levels and at various levels of granularity.

Also, the problem with current traffic flow prediction models is their inadaptability of detecting and tracking the traffic patterns changes [29]. There is a new pattern every time a non recurrent congestion occurs in the traffic flow and in this case, the model is not able to predict as accurately as when there is recurrent congestion. Existing approaches to traffic flow prediction do not adapt to the varying traffic situations because their distribution are memoryless, and they need a structure that will characterize the system at each step, not independently from the prior stage. To improve the flow prediction accuracy, a model should update from its normal path and track the changed traffic pattern, generating forecasts according to the new traffic pattern. Furthermore, at this time, operators necessitate extensive data sources to guarantee the accurate evaluation of the traffic state in real time because current traffic data collection systems do not incorporate the ability of registring detailed information on the altering events happening on the road, such as vehicle crashes, adverse weather, etc. Operators require external data sources to retireve this information in real time [71]. Besides, well-tailored data sources may not always be available for a particular area of the traffic network. Future systems should enable continuous monitoring of the traffic condition along all roads of the traffic network based on real-time information. Mainly, the problem is that most existing works ignore the context information when proposing models in their study of traffic flow prediction [92]. While using all context dimensions will provide the most refined information and thus lead to the best performance, it is equally important to investigate which feature or set of features is the most informative for the task of traffic flow

prediction on a segment. The benefits of revealing the most relevant context dimension include reduced cost due to context information retrieval and transmission, reduced algorithmic and computation complexity and targeted active traffic control.

With the advances in computers, which are more distributed, open, large, heterogeneous, and the progress in communication technologies such as cellular, satellite positioning and Connected Vehicles (CVs) technology enabling Vehicle-to-Infrastructure (V2I) and Vehicle-to-Vehicle (V2V) communications, transportation management is no more uniquely a civil engineering problem. In fact, connected vehicles evolve in a data-rich environment where they consistently generate and receive a variety of data [93]. In this article we show how integrating the transportation system with real-time information from connected vehicles to predict flows on target road segment results in a powerful tool for transportation analysis and evaluation. The ultimate goal is to build an intelligent transportation systems based on real-time information and for traffic management, the future will be in cooperative systems as they can benefit from the information collected from the vehicular ad hoc networks created by the connected vehicles. Self-organization is essentially a distinctive characteristic of CVs and represents a new approach in this domain as it is a new way of seeing transportation and their planning. Because traffic is essentially unstable, chaotic, far-from-equilibrium, and unpredictable, we propose a new data collection method in this domain with connected vehicles being the communication architecture for seamless exchange of information between the cooperative vehicles. A more precise view of traffic flows over the road network will be assessed so that traffic engineers can better layout a city's vehicular infrastructure.

Many commuters have an overall sense for the status of traffic and the overall times until congestion at bottlenecks will likely start and end, based on their long-term experiences. People may be familiar with typical traffic patterns, some situations whether traffic states of interest would be viewed as surprising. Current methods do not incorporate this overall sense, this experience, although, one of the hot topics in intelligent transportation systems (ITSs) is the development of distributed traffic information systems (TISs) [94]. Such distributed systems monitor and collect data from many sources. These data provide enough comprehensive information in order to better characterize the events detected. Current techniques fail to process the knowledge acquired from the data. With the widespread traditional traffic sensors and new emerging traffic sensor technologies, traffic data are exploding, and we have entered the era of big data transportation. In the big data era, techniques should be implemented to make use of the acquired information. Furthermore, at the current stage, the ITS is partially efficient since the vehicle as an entity is not fully contributing to the system. In fact, presently, vehicles are uninformative as they are not engaged in the process of traffic flow prediction. However, equipped with a communication technology, vehicles can exchange

information and cooperate collectively so as to provide their input to the system because of the unpredictable nature of traffic and because of the myriad factors that affect traffic flows such as weather conditions, the behaviour of other drivers, traffic issues, and other events.

This study presents a novel framework for the real-time distributed prediction of traffic flow in an urban network using connected vehicles technology. We foresee that an accurate prediction necessitates a mix of centralized and distributed system architectures through leveraging vehicular communication. Via this next generation sensing technology, we are interested in identifying road traffic events on the basis of exchanging traffic flow data between vehicles. If connected vehicles can detect congestion and cooperatively attribute a possible cause to it, we believe that they can then transfer this knowledge in real time to a central entity able to accurately predict flow on a road segment. Because the flow fluctuates from one time to another, it's better for a road side unit (RSU) to monitor the parameters for a period of time. Since some prediction techniques impose some constraints on the quality, type, and format of the used data feeds in order to ensure high level of accuracy, we see that the level of granularity given by connected vehicles will address this issue. We also focus on how the context information can be obtained from the exchange between vehicles while existing works ignore the context information. The basis of the prediction model lies in the fact that we integrate the impact of various events into the forecasting. Using historical flows and well engineered features, such as real-time reports from connected vehicles and travel time along a trajectory for accurate forecasting of flow in an urban network, the model learns a representation that takes into account the various events that vehicles realistically encounter on the segments along their trajectory. They may come across incidents, workzone, inclement weather, special events or recurrent congestion. All these situations are assessed by the connected vehicles and are represented by creative features to be fed to the model for the sake of learning to predict traffic flow.

We propose a Deep Neural Network (DNN), and tackle the problem by learning the target DNN in a multitask learning technique. DNNs have successfully been applied to traffic flow prediction, [67], [95]. One of the most important reasons for explaining their success in achieving state-of-the-art performance is their capacity to embody the characteristics of transportation systems, such as the large amounts of data and the high dimensions of features. Traffic data are exploding, and we have entered the era of big data in transportation and based on the fact that few things are very common but most things are quite rare, generalizing from relatively small sample sizes is still very relevant even in the big data era. This makes deep learning a promising method for transportation research.

Particularly, it was shown that it is possible to jointly train a model for solving different tasks

simultaneously, it is called multitask learning, MTL [62]. In machine learning, we normally break a complex problem down into tractable sub-problems, and learn to solve one problem at a time. This potentially ignores rich sources of information found in the training signals of other tasks. MTL is an inductive transfer between tasks. We conjecture that when the tasks involved in MTL are semantically connected, a larger improvement in predication accuracy can be obtained. More specifically, MTL can be more effective when we can encode the instances from different tasks using the same representation layer expressing similar semantics. Particularly, we show how traffic flow prediction falls into one domain of multitask learning. One domain is using the past and the future to predict the present. Often valuable features become available after predictions must be made. These features cannot be used as inputs because they will not be available at run time. If learning is done offline, however, they can be collected for the training set and used as extra MTL tasks. The predictions the learner makes for these extra tasks are ignored when the system is used ; their main function is to provide extra information to the learner during training. The valuable information contained in those future flow measurements help bias the model towards a hidden layer representation that better support flow prediction from the features that would be available at run time.

The results show our approach significantly outperforms the performance of the state-of-the-art ARIMA time series and baseline classifiers such as Random Forest (RF) and Artificial Neural Net (ANN). When compared with ARIMA and Random Forest, MTL presents an average performance in terms of root-mean-square error (RMSE) equal to 0.05. Compared to single task learning with ANN, our experiments with the data show that ANN have a lower performance (0.113 for RMSE) than MTL, but higher performances than ARIMA. This shows that when the tasks involved in multitask are semantically connected a larger improvement in accuracy of prediction can be obtained.

The contributions of this paper are summarised as follows :

— Monitoring of microscopic and macroscopic traffic variables via connected vehicles for the extraction of relevant contextual traffic features in order to summarise valuable knowledge in an efficient way.
— Forecasting of short term traffic flow on a target road segment with a Deep Neural Network trained to predict multi tasks with input from connected vehicles.
— Evaluation and validation of the proposed framework and inference method is made relying on simulation generated scenarios extended from a realistic data set of urban city vehicular motion traces.

This paper is organized as follows. After introducing the related work in Section 6.2, we describe our proposed framework in Section 6.3. In Section 6.4, we present the STP model.

In Section 6.5, we describe the simulation based on real vehicular mobility traces and provide analysis and discussion of the results. We conclude the paper in Section 6.6.

6.2 RELATED WORK

The traffic flow prediction problem aims at evaluating anticipated traffic flow at future time frames on a target road segment. One of the factors that affect prediction accuracy is data resolution which is highly dependent on the chosen forecasting horizon and step. The Highway Capacity manual [96], as well as some studies in the literature, have suggested 15 min as the most appropriate horizon value for short-term traffic flow prediction. For the step value, the most used value is a 5-min interval due to the high variability of the traffic flow. In general, traffic flow prediction techniques can be mainly classified in three categories : 1) parametric approach ; 2) nonparametric approach ; and 3) hybrid approach. In this section, we present the literature related to the prediction of traffic flow. We review it from two perspectives : the type of traffic data source used to collect the data and the technique used to model traffic as they are factors that affect the forecasting accuracy.

6.2.1 Parametric approach

The main techniques used in this category are time-series models, AutoRegressive Integrated Moving Average (ARIMA)-based models [46] and Kalman filtering [47]. In [48], they applied an ARIMA model for traffic volume prediction in urban arterial roads. Many variants of ARIMA were proposed to improve prediction accuracy, such as Kohonen-ARIMA (KARIMA) [49], ARIMA with explanatory variables (ARIMAX) [50], vector autoregressive moving average (ARMA) and space–time ARIMA [51], and seasonal ARIMA (SARIMA) [52]. Other types of time-series models were also used for traffic flow prediction such as the statistical models. They make the assumption of stationarity of the underlying process. This assumption is often violated as observable traffic conditions can evolve differently at different times. Also, the linearity of the time series approach presents an inconvenience for traffic prediction. Traffic flow has stochastic and nonlinear nature, unfortunately, even an enhanced ARIMA cannot accurately predict flow in the presence of accidents. ARIMA, due to its delayed reaction, is not an ideal method to use in the case of events which cause sudden changes in the time series data. If we know per say, from police event streams, that there is an accident (say, 30 minutes) ahead of us, we may be able to predict its delays and account for it. On the other hand, historical data can be used to identify similar accidents, i.e., with similar severity, similar location and during the similar time, so that we can use their impact on average speed changes and backlog to predict the behaviour of the accident in front of us. For

example, an accident that may happen between 4 :00PM and 8 :00PM on a particular road segment might cause 5.5 miles of average backlog ahead of the accident location. If the same accident happens between 8 :00PM and midnight the backlog will be 2.5 miles. In addition, these techniques predict traffic flow on each road segment separately. Since transportation networks are complex and much correlated, it is crucial to predict traffic flow from a network perspective. Moreover, while time-series analysis models are probabilistic, they are ignorant of the underlying process that generates the data. Thus, time-series-based approaches are more prone to large errors in traffic flow forecasting.

6.2.2 Nonparametric approach

Nonparametric regression [53] is a widely used technique. In [54], an online boosting regression technique that ensures traffic prediction under abnormal traffic conditions was proposed. Otherwise, boosting is disabled. In [55], a support vector regression was used to establish the prediction model, whereas particle swarm optimization was used to optimize the model's parameters. Among all of these techniques, neural-network-based forecasting had the best performance in terms of prediction accuracy and are considered to be relatively effective methods because of their well established models.

6.2.2.1 Neural Networks

A panoply of artificial neural networks (ANNs) were proposed to predict traffic flow [56]. Typical computational intelligence-based forecasting methods mainly include the back propagation (BP) neural network [53], radial basis function (RBF) neural network [57], recurrent neural network [58], time-delayed neural network [59], and resource allocated networks [60]. Particularly, deep learning is a neural network of more than one hidden layer. This technique has attracted researchers from various domains as it considers complex correlations between features and outputs. Besides the factors of scope, data resolution and technique used to model the traffic, we will compare the works in this approach with regards to the features used to train the models and the type of traffic data source used to collect the data.

In [61], they propose a stacked auto-encoder model to learn generic traffic flow features by considering the spatial and temporal correlations. The model is trained in a greedy layer-wise fashion. The traffic data are collected every 30s from over 15 000 individual detectors, which are deployed state wide in freeway systems across California. Their input features consists of traffic flow data at previous time intervals, on the target road segment. The target road segment is the link of interest in the road network where the model wants to predict flow on. Considering the temporal relationship of traffic, to predict the traffic flow at time interval

t, they use the traffic flow data at previous time intervals. In this study, their simulations indicate that four past time intervals of 15 minutes are enough to get good performance.

On another hand, recent work has shown that it is possible to jointly train a general system for solving different tasks simultaneously [62], MultiTask Learning (MTL). If the tasks can share what they learn, the learner may find it is easier to learn them together than in isolation. MTL is one way of achieving inductive transfer between tasks. The goal of inductive transfer is to leverage additional sources of information to improve the performance of learning on the main task. In [63] and [62], they train a MTL model to predict flows on links. Unlike traditional traffic flow forecasting that predicts a future flow of a certain link only using the historical data on the same link, which is also called single-link traffic flow forecasting, the authors propose multilink forecasting models, which take the relations between adjacent links into account. Single-link forecasting approaches ignore the relationships between the measured link and its adjacent links. In fact, each link is closely related to other links in the whole transportation system. The multilink model predicts traffic flows using historical traffic flow data from all of the adjacent links. The features in [63] are flow data collected from sensors on the road. In [64], they propose a combination of multitask learning and an ensemble learning method bagging, for traffic flow forecasting. In [62], they propose a deep architecture that consists of two parts, i.e., a deep belief network (DBN) at the bottom and a multitask regression layer at the top. In a transportation system, all roads and entrance–exit stations are connected to each other. There is a lot of shared information among these roads and stations. The data are collected from inductive loops continuously collecting data in real time for more than 8100 freeway locations throughout the State of California. In [65], they proposed approaches showing significant improvements in prediction accuracy when compared to baseline predictors but their focus lies on highways that are one-directional road segments, whereby usually in the inner cities the impact of traffic is a multi-dimensional problem, evolving in a 2D, more complex route network.

Current research on traffic flow prediction mainly focuses on data traffic history and neglects other conditions affecting traffic. In [66], they investigate and quantify the impact of weather on traffic prediction in a freeway scenario. They admit that transportation systems might be heavily affected by factors such as accidents and weather. But they just considered the weather factor. They claim that inclement weather conditions may have a drastic impact on travel time and traffic flow. Their MTL architecture incorporate deep belief networks for traffic and weather prediction and decision-level data fusion scheme to enhance prediction accuracy using weather conditions. The traffic flow predictions provided by their approach use past values of the traffic flow and the current weather data is fused to provide future traffic flow prediction. They state that their scheme avoids compounding prediction errors

that may ensue had weather data been predicted rather than been used as real information. Traffic flow is measured every 30 s using inductive loop sensors deployed throughout the freeways.

6.2.3 Hybrid approach

Some studies have investigated hybrid approaches [67]. To obtain adaptive models, some works explore hybrid methods by combining several techniques. Although the aforementioned hybrid models are flexible, they do not fully take profit from spatial information collected from the whole road network. Moreover, these studies rely only on information collected by sensors such as the Global Positioning System, loop detectors, and smart-phones. In traffic event analysis, the effect of events on traffic prediction has also been studied in the fields of data mining and transportation engineering. The majority of these studies focused on real time event/outlier detection using probabilistic or rule-based approaches (e.g.,[68], [69], [**?**]). There are also several studies that mainly concern the cause of the events, aiming at how to design the network or re-direct the traffic flows to avoid the delay of events [35]. However, none of these studies incorporate events into traffic flow prediction techniques, and hence fail to provide realistic forecasting in the presence of events.

In sum, the majority of the techniques focus on predicting traffic in typical scenarios (e.g., morning rush hours), and more recently in the presence of accidents or a weather condition. Existing techniques are only applicable to predict one of the scenarios. ARIMA prediction model is more effective in predicting the speed in normal conditions and not at the edges of the rush-hour time (i.e., the beginning and the end of rush hour). This becomes even more challenging when considering different causes for congestion, e.g., recurring (e.g., daily rush hours), occasional(e.g., weather conditions), unpredictable (e.g., accidents), and temporarily—for short-term (e.g., a basketball game) or long-term (e.g., road construction) congestions. These approaches consider traffic flow as a simple time-series data and ignore phenomenons that particularly happen to traffic data, the observations made in the immediate past are an indication of the short-term future. The statistical approaches, by their very nature the mathematics of collecting, organizing, and interpreting numerical data, can provide more insights on the mechanisms creating and processing the data. However, the statistical approaches frequently fail when dealing with complex and highly nonlinear data. Finally abnormal traffic patterns caused by non recurrent congestion or incidents may deteriorate the performance of these models [97]. Nevertheless, under most situations, extreme values are of primary interest in forecasting the change in traffic conditions. It is difficult to say that one approach is clearly superior over other approaches in any situation. One

reason for this is that the proposed techniques are developed with a small amount of separate specific traffic data, and the accuracy of traffic flow prediction methods is dependent on the traffic flow features embedded in the collected spatiotemporal traffic data. In general, literature shows promising results when using neural networks models as they are used as benchmarking methods for short-term traffic prediction [35].

The focus of our study is to integrate the impact of various events into forecasting models. The impact of events on traffic flow varies based on space and time. For example, the consequence of an accident occurring during rush hour is usually more severe. Similarly, an accident at an inter-state street has a different impact than that of a surface street. In this study, we consider such spatiotemporal characteristics of traffic in training our models because no studies have tackled the problem of analyzing the tight correlation between traffic data and external factors in an urban traffic network. It should be noted that the prediction of traffic flow under atypical conditions is evidently more challenging than doing so under typical conditions and, hence, much desired by operational agencies.

6.2.4 Design principle

A recent survey reported that the design principle of a prediction algorithm is to use a combination of historical data, real-time feeds, traffic modeling and simulation to predict how the traffic flow will evolve in the near future [1]. These steps will leverage properties of the road network such as the spatiotemporal correlation for faster inference. A prediction algorithm will use real-time traffic feeds i.e., traffic flow, travel time on a road segment or speed and apply an advanced modeling approach combined with historical data to predict the future traffic flow on a segment.

Historical data can be obtained from surveys, fixed monitoring equipments or real mobility traces, like GPS traces from floating car data for example, taxis as vehicular devices to collect GPS information. Taxi traces does not accurately represent origin and destination intentions of most travelers as they represent only one type of on road motor vehicle. Also, traces extracted from surveys or floating car data are not sufficient to record the statistical features of mobility in a large environment [11]. In contrast to these works, we use a realistic real-time motion simulation dataset to validate our work. We use a dataset that is representative in proportion to the city's vehicular mobility to extract historical data.

Real-time monitoring of traffic should be done at short intervals to provide good quality because stale data is useless in dynamic environments. In all the proposed methods in the literature, data is collected from fixed monitoring equipments, such as sensors and CCTV cameras, or using mobile data sources such as floating GPS data and SMS, social data feeds

[98]. The type of traffic data source is a very important factor as the heterogeneity of data sources and the variety of their format and level of granularity may add extra constraints on the designed prediction algorithm and may also affect its efficiency and accuracy. Even if Induction Loop Detectors (ILDs) are the most prevalent data collection technique and generally have the highest accuracy ; they can collect all of the fundamental traffic data except travel times. Some prediction techniques impose some constraints on the quality, type, and format of the used data feeds in order to ensure high level of accuracy. To the best of our knowledge, we are the first study to consider the connected vehicles technology for the collection of real-time feeds for the problem of traffic flow prediction on a road segment. We see that a much more efficient system would result if the vehicles of the connected vehicles themselves collect real time feeds because computation would aggregate a quality of data at a vehicle level instead of a quantity of data. Data are exchanged between connected vehicles every 0.1 seconds and technology on board of the vehicle will take both the macroscopic and microscopic level mobility parameters together into consideration in order to provide a robust framework.

Also, in all the proposed methods in the literature, the traffic data cannot be directly measured everywhere, but needs to be interpolated from incomplete, noisy and local traffic data at the specific location of the detectors, sensors or cameras. We want to maximize the coverage of the traffic. Connected vehicles are key players because they provide real-time traffic information along each step of their trajectory, hence allowing a reactive and dynamic traffic data estimation. This will help deduce more accurate traffic data especially over longer distances. Real-time urban monitoring through connected vehicles is proposed to obtain a global vision of the traffic in the network. Also, a significant challenge in terms of information gathering is related to the number of entities which collect traffic-based data, from road traffic operators such as public transport companies, private taxi companies, etc., and public traffic management authorities, to health and environment monitoring institutions, such as health boards, environmental protection agencies, etc., and private companies and individuals. All these data-gathering entities use independent measuring methods which acquire various data with different characteristics and using diverse methodologies and save it in their own databases. The most important consequence of this lack of a common format is the difficult synchronization of the information gathered by various sources, which makes the almost impossible coherent usage of information and cross correlation of events [89]. In the planning of our model, we make the most of the connected vehicles technology to collect both traffic and event data. We utilize whitin the same gathering entity, connected vehicles that is, optimal strategies for the monitoring of data and collection in order to enhance the efficiency and thus the performance of the system. A portion of our model is built offline by using the

historical data to assess the mean traffic variables of a road segment, we then use it online for short-term traffic flow prediction. In real-time using the current reports from connected vehicles as input to predict flows. Also, traffic events are assessed in real-time again by the connected vehicles and are given to the predictor online for better traffic flow prediction.

We present in the next section our framework. We focused on defining real-time forward looking analysis techniques that use historical traffic data and real-time traffic feeds more fitting precisely a more complex urban road network than the freeway network to predict how traffic flow will evolve on a target road segment.

6.3 FRAMEWORK

The framework proposed in this study is semi-centralized. On one hand, connected vehicles collect and propagate data via the ad hoc networks formed between them along a route. On another hand, a road side unit (RSU) is installed on a target road segment and collects data for a period of time to get a clearer picture about the traffic on the road. This spatio-temporal collection of traffic information by both entities leads to a more scalable structure and is described in the following sections.

6.3.1 Data collection by CVs

Using vehicle-to-vehicle communications, a connected vehicle can continuously collect traffic characteristics representing the evolution of traffic over time on the road network. Consequently, a huge amount of traffic data can be archived at a vehicle level. We propose the collection of real-time traffic data from connected vehicles because to a large extent, the quality of observed traffic data will affect the accuracy of prediction. Vehicles use broadcasting as a data forwarding, allowing data to move faster than the speed of traffic. In our design, it is not required for vehicular networks to remain continuously connected (the network can temporarily split). In fact, a vehicle should be able to enter story-carry-forward mode if there are no vehicles in his vicinity. For example a vehicle can take 200ms to deliver a message 4 km away under normal traffic conditions. It may take less than 200ms under high traffic congestion. It may also take 60s to deliver a message 4km away in a story-carry-forward way.

Upon investigation of data on board of the vehicle, each vehicle computes travel time on each segment of its trajectory. Travel time is the key data in this study. Travel time is the main factor to affect traffic flow, almost all other factors are closely related with the travel time, and it has the same trend to rise and fall. Induction loop detectors are the most prevalent, generally have the highest accuracy, when collecting all of the fundamental traffic

data (density, time mean speed, space mean speed, etc.) except for travel times. Although with connected vehicles, it's not easy to collect density because this metric is affected by low message reception rate in case of network disconnectivity or low penetration rates, connected vehicles can straightforwardly collect travel times along each segment of their trajectory. They can also collect those of others by cooperation between them. Then, each vehicle recurrently confronts its local network view with the other views so as to update it. The vehicle then broadcasts to its surrounding its traffic data. The vehicle collects assessments from others in a table. This propagation process is shown in Fig. 6.1.

Each time a vehicle receives travel time information broadcasted by another vehicle, it updates its stored data accordingly, and the vehicle has to pass on the travel time data sent by the neighbours to other neighbours in its coverage area. To do so without flooding the network, each vehicle firstly computes a TravelTime Index of their own and secondly they average travel time of others. The index is representative of the observed travel time trajectory of the vehicle compared to historical travel time values along the trajectory. It is the sum of the weighted average of the difference in travel time on a link, for all previous segments of the trajectory. The weights in the index around the current segment increase so to better capture current local view. The equation is as follows :

$$TTindex = \frac{\sum_{i=1}^{10}(1 - exp^{-\frac{i}{10}(\frac{TTi-TTh}{TTh})}}{10} \tag{6.1}$$

After computing the TravelTime Index of their own, vehicles have to average travel time of others. In order to reduce randomness, the average method is taken to calculate the travel time and considering the impact of the travel times of all vehicles within the scope, the formula is set as follows :

$$v = (1 - \alpha) * vs + (\alpha) * vr \tag{6.2}$$

where vs is selected as the travel time index of the vehicle, vr is a mean value of the vehicular travel time indexes in the wireless coverage of the vehicle and α is a weighting factor which means the different degrees of importance. After experimentation we fixed α to be 0.65.

Also, connected vehicles support delivery of vehicle-to-vehicle messages that are sent every 0.1 seconds. Therefore, networked cars can be extremely fast in warning their surroundings regarding a blocked road, accident, a special event, etc. The local recognition of this type of road traffic information is equivalent to the investigation of real time outlier in the traffic flow series. Outliers represent potentially extraordinary patterns in traffic flow series. It

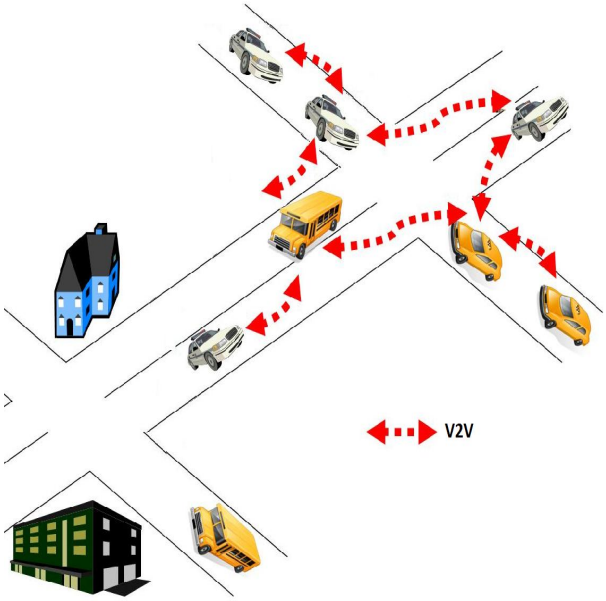

Figure 6.1 Propagation process of the connected vehicles to collect data

was recently proved that if the outliers can be detected, then they should be assimilated into the forecasting system for adaptively responding to the changing traffic patterns. In fact, connected vehicles can detect extraordinary patterns and hence supply more insight for the short-term traffic flow forecasting system. The architecture incorporates the ability of registring in vehicle detailed information on the transient altering events along a vehicle's trajectory, such as vehicle crashes, adverse weather, etc.

To do so, we firstly implement on board of each vehicle the algorithm for the detection of congestion via connected vehicles presented in [19]. A congested segment is a piece of road network where the difference between expected and actual travel time is bigger than a certain threshold. The expected travel time on a given segment can be obtained from a number of historical observations. In this way it is possible to learn the typical traffic behaviour for a given road segment on a certain day of the week, at a given time. If a segment is congested, it takes more time to pass by it. Understanding the correlation between traffic indicators and road traffic events is crucial. A rigorous analysis was done in [83] and proved that connected vehicles are able to collect traffic data (travel time, gap, speed, problematic spots) and estimate the event on the segment when excessive congestion is experienced. Excessive congestion considers the additional time taken to pass through a road segment in comparison to a threshold. Thus, we also implement on board of each vehicle the algorithm in [83] that

permits vehicles not only to detect congestion but also classify the cause of congestion.

In fact, traffic bottlenecks are disruption of traffic and are of two general types stationary and moving bottlenecks. Stationary bottlenecks are those that arise due to a disturbance that occurs due to a stationary situation like narrowing of a roadway, an accident. Moving bottlenecks on the other hand are those vehicles or vehicle behaviour that causes the disruption in the vehicles which are upstream of the vehicle. Moving bottlenecks are caused due to slow moving vehicles that cause disruption in traffic. Moving bottlenecks can be active or inactive bottlenecks. If the reduced capacity caused due to a moving bottleneck is greater than the actual capacity downstream of the vehicle, then this bottleneck is said to be an active bottleneck. Bottlenecks are important considerations because they impact the flow in traffic. Suppose that, at time t, a truck on a road segment slows from free-flow to v. A queue builds behind the truck. Within the region of the truck, vehicles drive slower and a queue will back up behind the truck and eventually crowd out the street. The real-time monitoring done by the connected vehicles is local and self-organized and the results of the locally observed traffic situation are disseminated reactively. The model adaptability of detecting and tracking the traffic patterns changes by integrating the real-time feeds results will learn to differentiate between temporary induced traffic pattern change that is mitigated in a short period vs permanent pattern change in order to ensure a robust forecasting system.

6.3.2 Data collection by RSU

On the target road segment where traffic flow prediction is required, an RSU is installed and continuously computes and stores current flows on the segment. It also collects specific data from vehicles passing by. Fig. 6.2 shows the deployment of an RSU on a target road segment.

An RSU can have more complete knowledge of its realm, assuming it can receive and hold information originating from anywhere on the map. When the RSU launches a request for prediction, all vehicles on the segment are involved in the process. In this manner, the data they transmit to the RSU is no longer a local static segment traffic flow data because it is not limited to the vehicle's assessment but to that of all the vehicles each vehicle encountered along its trajectory. To limit the analysis to a specific geographical area, because flow on a road segment is correlated to its surrounding and on average a node has an area of interest of 2Km, we impose the vehicles compute data from the last 10 segments of their trajectory. Also, the collect would also have to be restricted in terms of duration since values may change and new vehicles may appear in the area and contribute to the collect with new values. The maximal duration used by each vehicle is given by the successive values in its table for no more than 15 minutes. The RSU collects from CVs, TTindex, flows and events on other

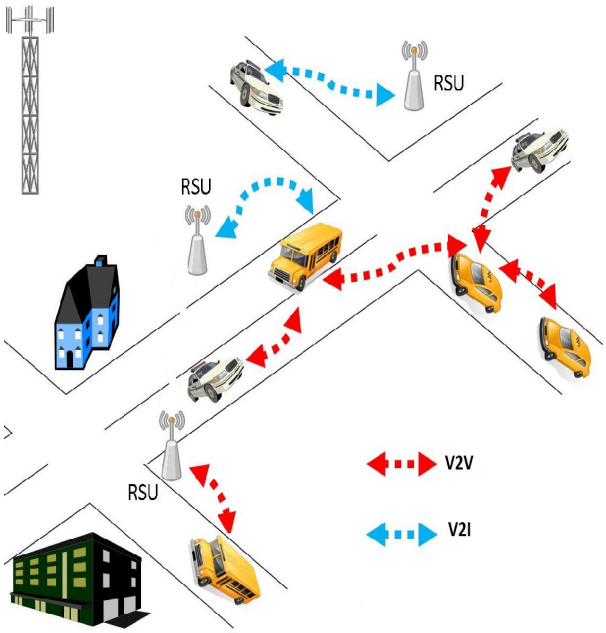

Figure 6.2 Deployment of a RSU on the target segment

segments and does the mapping between all influenced road segments. Around the target segment, the influenced road segments are those whose flow readings show an anomalous decline compared with the historical average flow when an event occurs in the surrounding. The idea is that real-time data (time, currents flows/events on adjacent segments and travel time experience around the target segment) and past traffic flows on the target segment are explanatory variables for prediction of traffic flow on a segment. The purpose is to use these information to predict traffic flow by means of a deep learning technique that learned from an adequate dataset to automatically infer from the correlations between these variables. The predictor takes input of the features sent to the RSU, and output the predicted flow in the predetermined future on the target segment. The predictor is the STP model and is presented in the next section.

6.4 STP Model

The traffic flow prediction problem can be stated as follows. Let $Xi(t)$ denote the observed traffic flow quantity during the tth time interval at the ith observation location in a transportation network. Given a sequence of observed traffic flow data, $i = 1, 2, \ldots, m$, and $t = 1, 2, \ldots, T$, the problem is to predict the traffic flow at time interval (t+Δ) for some

prediction horizon Δ. This is the short-term traffic flow prediction problem, STP. Some other works focus on predicting the traffic flow of the next several time intervals from $T + \Delta + m$ to $T + n$ as well, it is called the long-term traffic prediction. In our work, we consider the short-term traffic flow prediction problem.

On the other hand, most models in the literature predict flow $Xi\ (t+\Delta)$ at time $(t+\Delta)$ based on the traffic flow sequence $X = \{Xi,\ t/i\ \epsilon\ O,\ t = 1, 2, \ldots, T\}$ in the past, where O is the full set of observation points (roads and stations). The problem becomes, given the feature X and task Y pairs obtained from history traffic flow $\{(X1, Y1), (X2, Y2), \ldots, (Xn, Yn)\}$, learn the best parameters for a predicting model that minimizes a loss function. This is supervised learning because each input can be tagged with the flow Y corresponding to the next value in the time series obtained offline. However, in our work, we consider that traffic flow prediction not only depends on historical flow data but heavily depends on real-time traffic data. To do so, we incorporate to the input feature X, not only previous traffic flows observed on the target road segment but knowledge acquired from related roads. The model is fed well engineered features, such as real-time reports from connected vehicles and travel time along a trajectory in order to learn a representation that takes into account the various events that vehicles realistically encounter on the segments along their trajectory. As with most data mining problems, when modeling real-world physical systems, having good features is critical. In this study, the features are : current time of day, observed travel time trajectory of vehicles around the target segment (TTindex), past successive flow values on the target segment, flows on links around the segment and the presence of any traffic event on surrounding segments, such as incident, weather, special event, workzone or recurrent traffic. In order to store and exchange traffic characteristics, each vehicle creates information structures and stores them in vehicle. The structures consists of the vehicle's own measurements and measurements obtained from others. Each structure consists of the following fields :

- Timestamp : time of the measurement's creation to ensure freshness of measurements and to prioritize most recent ones.
- TravelTime Index : the observed travel time trajectory of vehicles around the target segment.
- Historical flow table : Past four flow values on the target segment.
- Adjacency flow table : flows on each of the eight neighboring segments.
- Adjacency event table : any traffic event on the eight segments around the target segment.

Particularly, in this study, the problem of predicting short-term flow is handled as a classification task. In fact, we propose that the target variable Y represent multiple classes of

discrete interval of flows and the task is for the classifier to predict the range of flow that the current traffic situation will generate at a near future time.

Moreover, we propose that the classifier learns to solve multiple tasks at the same time. In machine learning, we normally break a complex problem down into tractable sub-problems, and learn to solve one problem at a time. This potentially ignores rich sources of information found in the training signals of other tasks. It is possible to jointly train a general model for solving different tasks simultaneously. The classifier will prefer hypotheses that explain more than one task, improving generalisation. In [78], they proposed multitask learning (MTL) as a means of inductive transfer between tasks. The update is done with error signals of other tasks. Precisely, to use MTL for time series prediction, we use a single model with multiple outputs, each output corresponding to the same task at a different time. If output k referred to the prediction for the time series task at time *Tk*, the model makes predictions for the same task at three different times. The output used for short-term flow prediction would be the middle one so that there are tasks earlier and later that the model trained on. In particular, we propose that given a fresh new road network traffic situation at time t, *Xt*, the first task consists in determining what flow $c \epsilon Y$ is a suitable flow prediction at *t+5*. The second task is to find what flow $c \epsilon Y$ is a suitable short-term flow prediction at *t+15* based on the similar road network traffic situation and on the relevant prediction of the first task and the third task is to find the flow at *t+20*. Fig. 6.3 presents a road network traffic situation at different time periods. In the figure, the traffic is monitored on segments 1-8. The target segment is number 8. There is an accident on segment 7 at time t. The flow on segments one to eight is monitored at *t+5, t+15* and *t+20*. At time t, the first task is to predict flow on segment 8 at time *t+5*, the second task is to predict flow at *t+15* and the last task is to predict flow at *t+20*.

Given a fresh new traffic situation on a target road segment, we propose a feedforward neural network or Multi-Layer Perceptron (MLP), that solves the three tasks. The MLP is a series of logistic regression models stacked on top of each other, with the final layer being another logistic regression because we are solving a classification problem. The purpose of the hidden units is to learn non-linear combinations of the original inputs. Also, a transportation system is a highly correlated network. The characteristics of transportation systems, such as the large amounts of data and the high dimensions of features, makes deep learning a promising method for transportation research.

We extend the MLP to a Deep Neural Network (DNN), and tackle the problem by learning the target DNN in a multitask learning technique. We can easily extend the model to predict multiple outputs in order to do multitask learning. We conjecture that when the tasks invol-

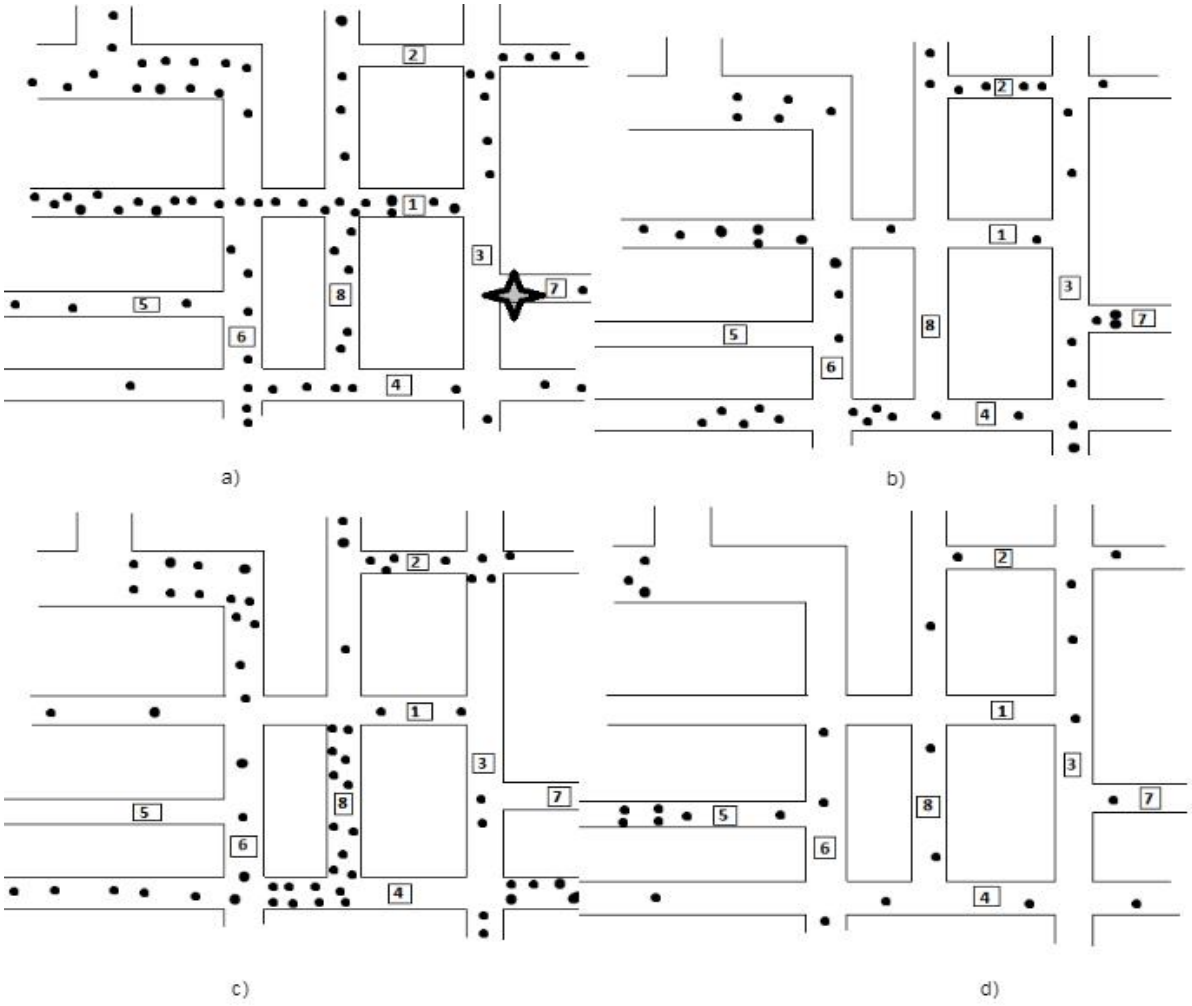

Figure 6.3 Monitoring of traffic on segments 1-8 at time a)t, b) t+5, c) t+15 and, d) t+20

ved in MTL are semantically connected, a larger improvement in predication accuracy can be obtained. More specifically, MTL can be more effective when we can encode the instances from different tasks using the same representation layer expressing similar semantics. Using historical flows and well engineered features, such as real-time reports from connected vehicles and travel time along a trajectory for accurate forecasting of flow in an urban network, the model learns a representation that takes into account the various events that vehicles realistically encounter on the segments along their trajectory. They may come across incidents, workzone, inclement weather, special events or recurrent congestion. All these situations are assessed by the connected vehicles and are modeled by creative features to be fed to the DDN for the sake of learning to predict traffic flow. The supervised multi-task learning DNN

model is presented in Fig. 6.4.

The input of dimension 62 in our joint learning architecture feeds three hidden layers. The supervised classifier has 20, 40 and 20 hidden units in the different layers. Three outputs are fully connected to the hidden layer that they share. Each output of the network contains four neurons representing the class label. Thus, the multi-task model is trained for classification on labeled examples. If we add mutual inhibition arcs between the output units, ensuring that only one of them turns on, we can enforce a sum-to-one constraint, which can be used for multi-class classification. Multi-classification means that target variable represents whether a traffic situation will generate one of four ranges of flow. Each output node estimates a conditioned class posterior probability given an input feature vector. The output is passed to three independent softmax to produce the scores for the individual tasks. By assigning a softmax activation function, a generalization of the logistic function, on the output layer of the neural network, the outputs can be interpreted as posterior probabilities. This is very useful in classification as it gives a certainty measure on classifications. The softmax activation function is :

$$yi = \frac{e^{xi}}{\sum_{j=1}^{c} e^{xj}}$$

Given the strong connection between the objective functions of the DNN, training is performed equally for all tasks using backpropagation. The function to be optimized is the mean squared error between network outputs and targets. The model learns the best parameters for predicting $\widehat{Y}$ that minimizes the loss function, i.e.,

$$L(Y, \widehat{Y}) = \frac{1}{2}(Y - \widehat{Y})^2$$

Models require the availability of a dataset of training and ground-truth annotations for classification. The model's accuracy strongly depends on the amount of training data and the variation within it. We present in the next section the simulation outline that help create the synthetic dataset and we provide results.

6.5 SIMULATION AND RESULTS

Economic issues, lack of large scale deployment and technology limitations make theoretical analysis and simulation the main choices in the validation of VANET. The realism of the simulation is thus a paramount aspect. Our experiments utilize a validated real-world traffic dataset of the City of Cologne, Travel and Activity PAtterns Simulation (TAPAS) Cologne

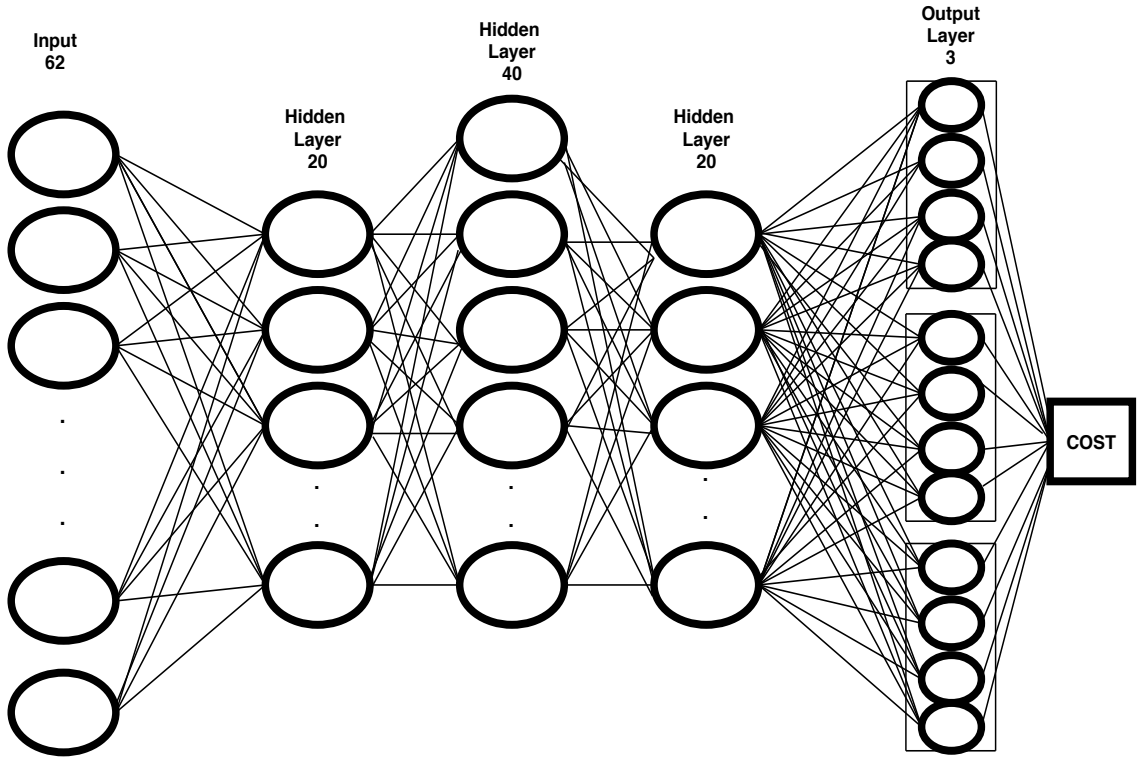

Figure 6.4 Multi-task learning DNN

scenario, assumed to be one of the largest traffic simulation data set [45].

6.5.1 Simulation outline

TAPAS covers the main road network within the inner city of Cologne. In the scenario, traces for the 6-8am peak hours are provided. We create extended scenarios mounted on top of the base scenario to model atypical traffic conditions such as weather, incident, workzone, special event and recurrent congestion. Our experiments are then built on the extended scenarios. Evaluation of our framework using complex real-world scenarios allow determining whether the proposed models can handle the real life's complexity. We use SUMO, a microscopic traffic simulator for the simulation of urban mobility [72]. Our investigations need the microscopic view for different reasons. Fine microscopic simulations model each vehicle explicitly and compute the traffic flow's progression by modelling each vehicle's speed and lane choice, mostly using discrete time steps of one second, calculating different traffic specific values like the amount of vehicle in a specific point and so on. Also, simulating a large area is necessary in each scenario because trajectory data along a vehicles' trip needs to be collected. Finally, SUMO enables generation of trace files that are necessary for the simulation of communication

in a VANET in the network simulator ns-2 [76].

In a simulation, atypical traffic conditions are not direct model parameters but must be converted into ones. We describe below the extended scenarios of atypical traffic conditions simulated using SUMO :

1. Extended Scenario of an incident : On the base scenario, we stop two or three vehicles, for a specific amount of time, on a lane to simulate incidents. We simulate incidents at the beginning, middle and end of a lane. We also simulate incidents on different lanes, for a long or short duration as well as inside or outside of an impact region of a special event.
2. Extended Scenario of a workzone : Similar to the above extended scenario, we stop vehicles on an edge to simulate a workzone. We vary the position on the edge and the duration.
3. Extended Scenario of bad weather : We convert the base scenario into an extended scenario of bad weather, snow for example. Snow might lead to slippery roads and reduced sight, leading to decreases in the vehicles' velocities and a more careful and defensive driver behaviour. Such behavioural changes would be reflected in simulation parameters, such as the driver's preferred velocities. Parameters of the car-following model are affected by the weather.
4. Extended Scenario of a special event : To generate trips to a particular destination edge where there is a special event, we have to generate random departures and random routes. We use a Poisson process to generate random timings for trips. Departures will occur individually, stochastically independent to all the others in the road network, at random moments. The rate parameter λ is the demand per second from different sources in the network, and can be seen as the flow. To generate random routes, given trips are assigned to respective fastest routes according to their departure times and a given travel-time updating interval by SUMO's traffic assignment model.

We then perform experiments on these extended scenarios based on real-world traffic. We assume that vehicles are equipped with a Global Positioning System (GPS) device for positioning and a detailed enriched digital road map for route guidance including the length of each road, number of lanes per road, the coordinates of the markers on the road and historical threshold values of travel time and flows for each road derived offline using past historical data. Also, vehicles are equipped with a transceiver for communication using Dedicated Short-Range Communications (DSRC) IEEE 802.11p. We use well-known neighbor discovery, geographical routing, and message forwarding techniques to pass tasks and information through the connecte vehicles. When vehicles on the congested segment experiencing

congestion detect that the observed travel time is excessive, the observable trajectory characteristics and the results of the local traffic evaluation are collected. We generate urban mobility traces from the extended scenarios for usage in ns-2, the discreet-event network simulator, in order to carry on the cooperative process [76]. Data of independent vehicles passing on the target road segment are collected. Characteristics are extracted from several scenarios, experiments and vehicles respectively and put into supervised feature vectors. We construct a synthetic training dataset. Synthetic datasets are designed to obtain information while still maintaining statistical properties of the original data.

On the target road segment, traffic flow is measured every second but is aggregated into 5-min duration. All inputs of the prediction model are real numbers except for the traffic events happening on the segments. Traffic events consist of one of the six causes : $\langle$accident, workzone, weather condition, recurrent, special event or no events $\rangle$. This order of the causes is important because the values representing events are given to the model via an encoding rather than as simple continuous inputs. The encoding we use is a one-hot vector of six values either being 0 or 1 in the presence of the event in the previous order. For instance, in presence of an accident on a segment, the one-hot vector is $\langle 100000 \rangle$and in presence of a workzone, the vector is $\langle 010000 \rangle$.

Once we have obtained the features, we map the prediction variables to classes or bins. There are many ways to construct bins, we experimented with many of these approaches and chose to do manual binning. This allows us to divide the range of values into sub-ranges. Once we have constructed the feature vector and mapped the continuous prediction variables to discrete classes, we train the STP model. The model will learn historical similarities between road segments. In contrast to using direct historical trends in the measure itself, this is more powerful since sometimes trends may not exist in the measure but do in the similarities.

6.5.2 Results

Fig. 6.5 shows the structure of demand data over 24 hour period in the whole region. This shows normal behaviour of mobility where peak hours like 6-9am, 4-6pm during which the number of trips are usually high. In our study, flows from 6 :00 am to 8 :00 am are collected on a target road segment under different scenarios. In our study, for a period of five minutes, for every 4-period window, a feature is created by incorporating the 5th, 6th and 7th data points as tasks 1, 2 and 3 on a target road segment. The next feature includes the 5th data point as input and task 3 becomes the 8th data point. This process continues until the last observation of flow at 8 :00 am.

For illustration purposes, we plot in Fig. 6.6 the profile of traffic flow values for a signalised

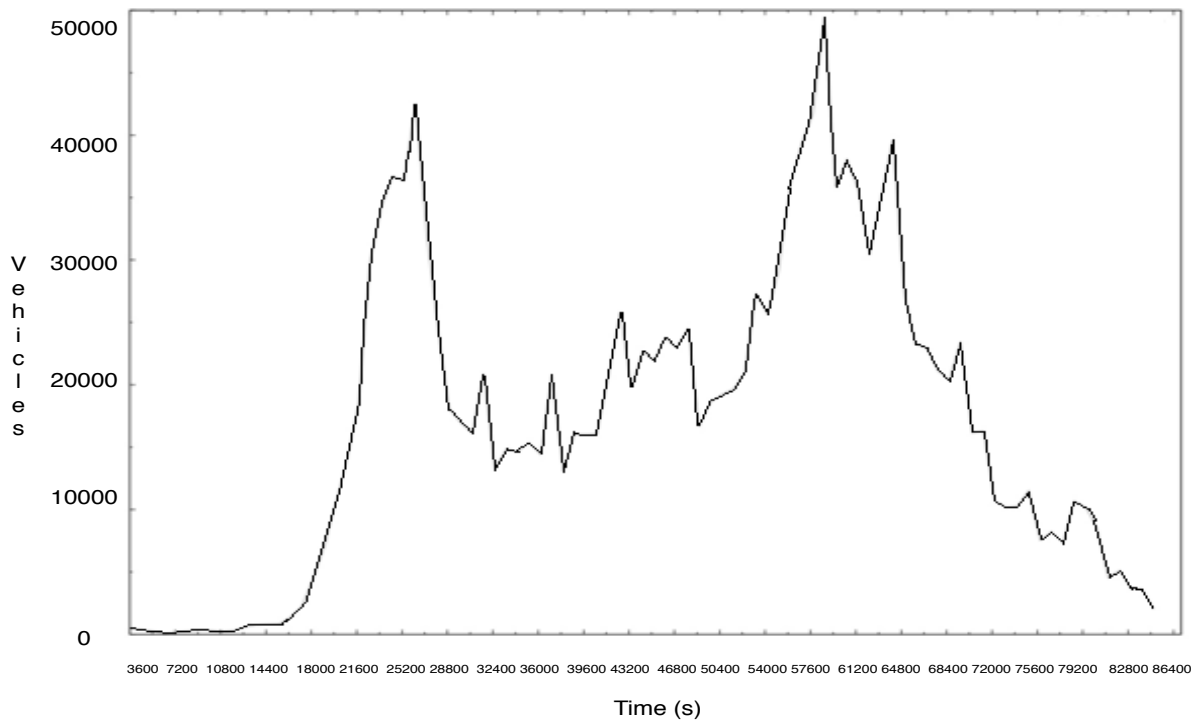

Figure 6.5 Demand data over a 24 hour period

road segment in the urban network. From this figure, we can sense that vehicles stop at the light and then another wave starts periodically. This behaviour of traffic flow is different from that occurring on a freeway. In the figure, no congestion and no events are present on the road segment. In Fig. 6.7 however, we show the profile of traffic flow values in the advent of an incident occurring on the urban road segment. We notice as congestion installs, how the flow values stay very low because density is high. We demonstrate how our model accurately predicts future flows in presence of any cause of congestion.

To make the proposed framework tractable, we compare the performance of our MTL model with various prediction models. Firstly, since our multitask learning model is built on MLP network, it is worth to compare and investigate how much improvement we could achieve beyond the baseline MLP network classifier. In Fig. 6.8, Fig. 6.9 and Fig. 6.10, we present three net architectures. In Fig. 6.8, MLPa is a standard net that learns the task of short term traffic flow prediction 15 minutes later. In Fig. 6.9, MTLa is a net that learns two tasks of prediction of traffic flow with the first task being prediction of flow 5 minutes later, that is before the target time and the second task is the main task of prediction of short term traffic

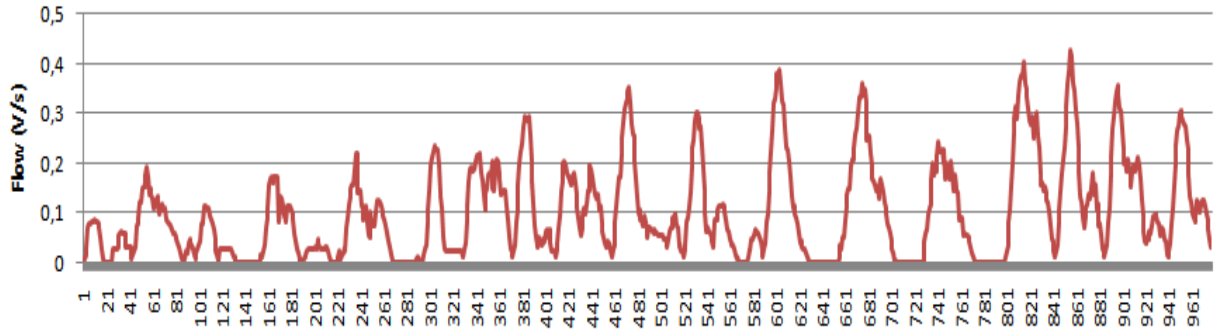

Figure 6.6 Profile of traffic flow on a signalised road segment

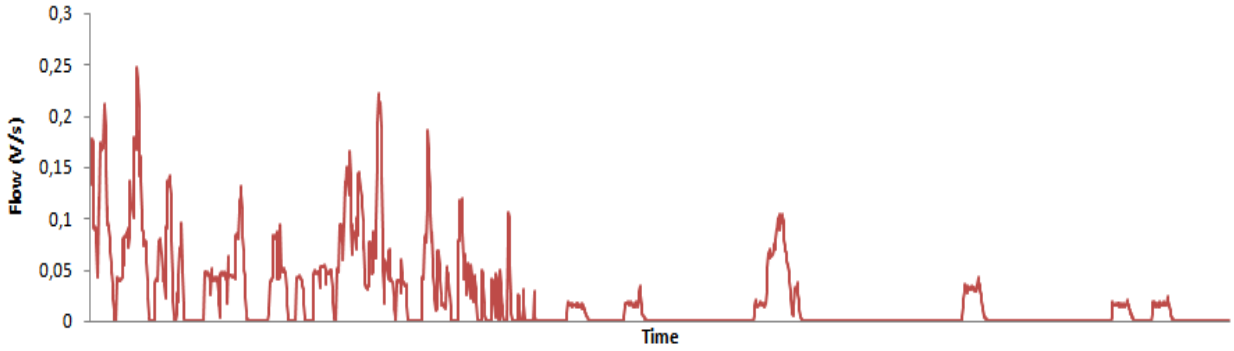

Figure 6.7 Profile of traffic flow on a signalised road segment in advent on an incident

flow 15 minutes ahead. In Fig. 6.10, MTLb is a net that learns two tasks of prediction of traffic flow with the first task being the main task of prediction of short term traffic flow 15 minutes later and the second task is prediction of flow 20 minutes later.

Also, to measure the predictive power of the proposed MTL model, we compared it with the performance of the state-of-the-art ARIMA time series approach and with a baseline classifier, Random Forest (RF), implemented in Weka. MTL and MLP models are implemented using Torch 5 package. Specifically, when evaluating the performance of our model, we use root-mean-square error, MSE. We use it to measure the linear score that averages the error with the same weight and to measure the residuals by assigning larger weights to larger errors.

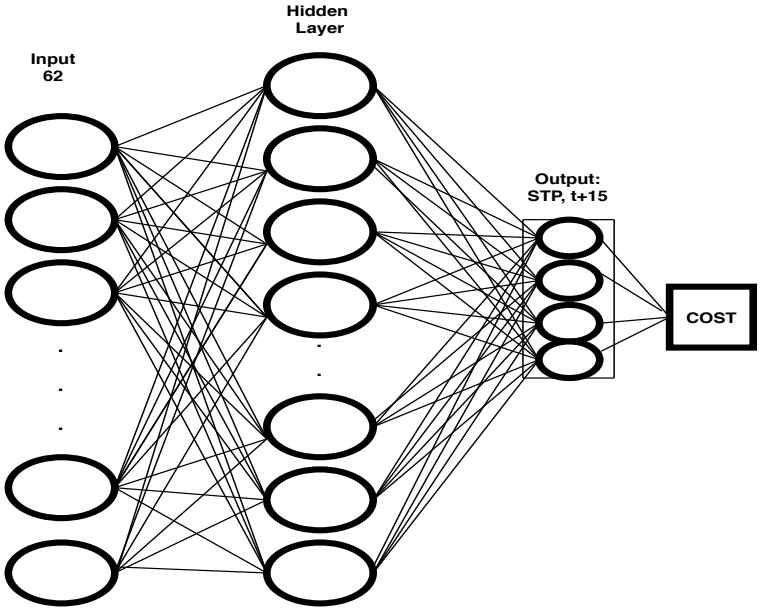

Figure 6.8 MLPa is a standard net that learn STP.

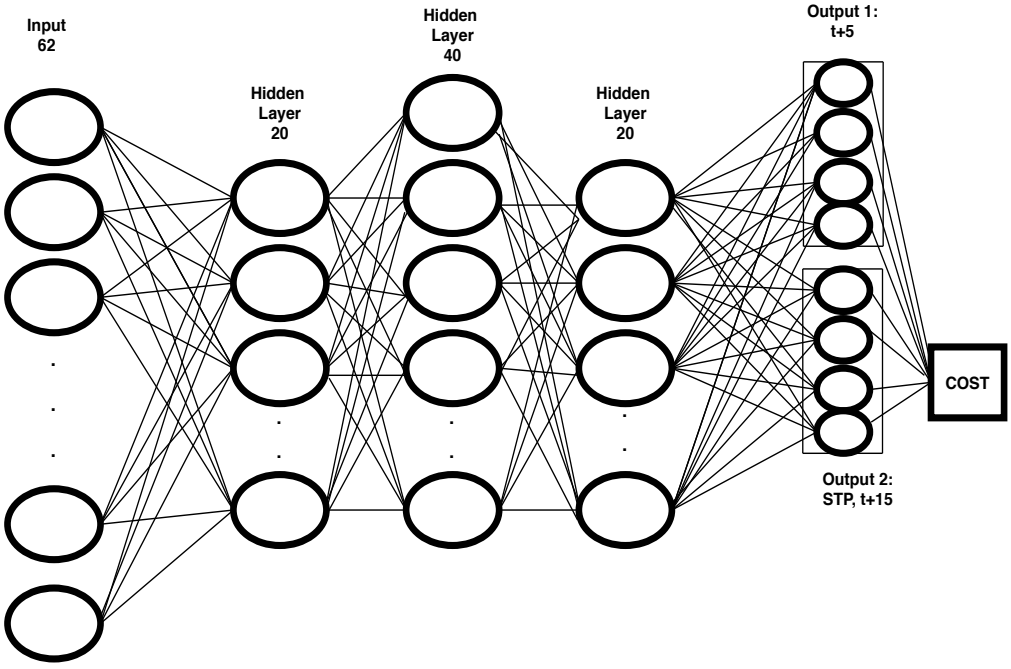

Figure 6.9 MTLa learns STP and flow at t+5.

$$MSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(Yt - \widehat{Yt})^2}$$

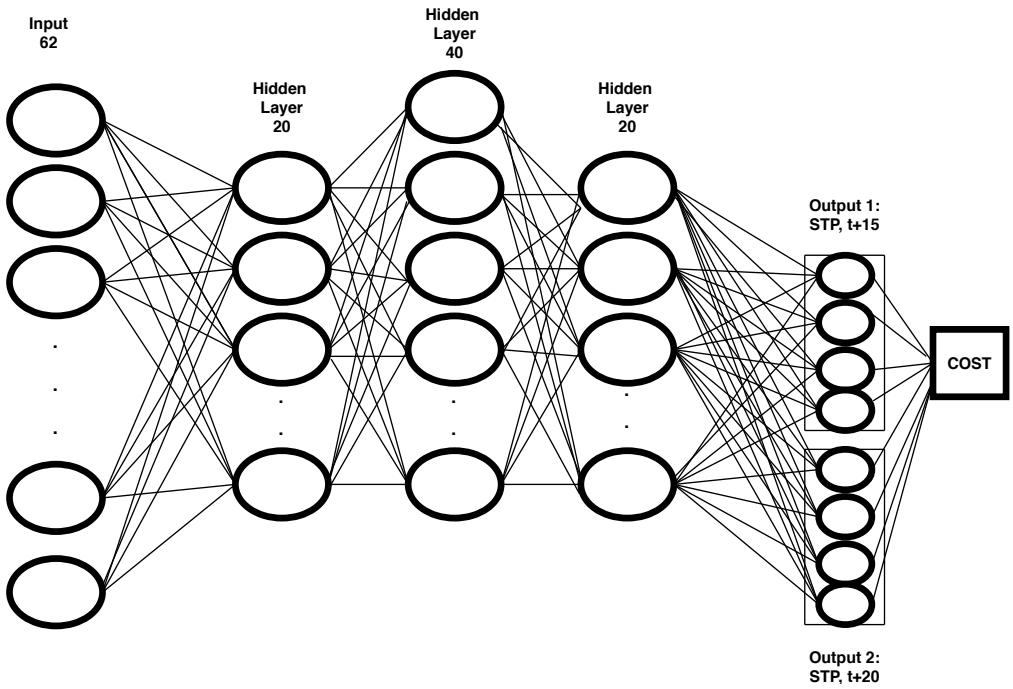

Figure 6.10 MTLb learns STP and flow at t+20.

We feed ARIMA the original traffic flow data. We implement ARIMA starting with stationary verification, followed by the iterations of 1 to 10 for Auto Regressive model and 1 to 10 for Moving Average model to reach the best combination under Bayesian information criteria. We use the trained model for one-step forecasting. We iterate the prediction procedure four times by using predicted value as previously observed value. For RF and the neural networks however, the data is not only traffic flows but also the other features proposed in this paper. In fact, the data for the above-mentioned tasks contains 6938 traffic situations related to five different scenarios. The dataset is then divided in training, validation and test sets. For the one-hidden-layer MLPa, we set the number of hidden units and epochs through cross-validation on the training data. The hidden units are varied between [5, 150] in steps of 5. Moreover, the number of epochs is varied between [25, 250] in steps of 25. We added dropout between all the layers of the network to improve generalization and avoid co-adaptation of features. We tested different dropout rates (0.2, 0.4) for the input and (0.3, 0.5, 0.7) the hidden layers obtaining better results with highest values, i.e., 0.4 and 0.7. The best architecture for MLPa is one hidden layer with 90 hidden units, and the number of epochs is equal to 150.

On the other hand, for the deep architectures, there are more parameters that we have to define, such as the nodes in each layer, the layer size and the epochs. These parameters are also determined through cross validation only on the training set to ensure fairness when

comparing with other approaches. The experimental evaluation is based on five folds cross-validation (CV) with 20 randomly repeated CV runs to obtain average performance scores for comparisons. Backpropagation is done on all outputs. To avoid overfitting, we did not try very deep MLP architecture. In the case of MTL, MTLa and MTLb we choose the layer size from two to five layers. The number of nodes in each layer is chosen from [5, 150] in the steps of 5. The number of epochs is crucial to the learning phase. Therefore, we vary the epochs' range from 50 to 300 in the steps of 50. We first randomly choose each parameter from the possible set, and then, we choose the best configuration from the random runs. The best architecture for MTL consists of three hidden layers with 20, 40, and 20 hidden units in the first, second, and third hidden layers, respectively. The best number of epochs of the MTL training is found to be 100 epochs. Since we only employ the synthetic dataset for training, models with very complex structures would be underfitted. More nodes in each layer would cause unnecessary burdens for model training and compromise the performance.

Table 6.1 shows the results of our MTL model in comparison with the time series, baselines and MTLa and MTLb using MSE. The scores are averaged from 20 randomly repeated 5-folds CV runs. MLP model makes comparable performance to the state-of-the-art RF model. But RF and MLP achieve better performance than the ARIMA time series, with error values of 0.122 and 0.113 respectively. This is expected because of the added features in the input vector. Also, deep networks perform better than baseline MLP because deep networks learn sub-features in the different layers to better characterise the output flow. Consequently, in all scenarios of traffic congestion due to different event, deep networks track better the sudden flow changes and their pattern. Particularly, results indicate that multi-tasking improve the performance compared to single task learning with MLP. Task 2 in MTLa and task 1 in MTLb try to capture the information contained in the training signals of other tasks drawn from the same domain. The tasks in these models exploit the joint input. If the tasks can share what they learn, the model performs better when it learns them together than in isolation. The difference between MTLa and MTLb is in the training phase. Because of the joint representation, MTLa is 0.04 better than MLP on the test set result and MTLb is about 0.028 better than MLP.

We analyze the contribution of MTL to the prediction problem. We notice that some hidden units of MTLa and MTLb became specialized for just one or a few tasks. Task 2 in MTLa and Task 1 in MTLb need to compute the same subfeatures. If Task1 from MTLa and Task 3 from MTLb are used as extra outputs in MTL, this signify that they must be learned ; it will bias the shared hidden layer to learn the input features better, and this will help the MTL net better learn to predict outputs. This confirms the importance of having highly related tasks and our idea of using MTL to improve the target Task 2. MTL provides the best MSE,

Table 6.1 Performance comparison of MTL with the time series, baselines (RF, MLP) and MTLa and MTLb using MSE.

Task	ARIMA	RF	MLP	MTLa	MTLb	MTL
5-min Traffic Flow prediction	-	-	-	0.042	-	0.056
15-min Traffic Flow prediction	0.255	0.122	0.113	0.073	0.085	0.052
20-min Traffic Flow prediction	-	-	-	-	0.094	0.108

improving MTLa and MTLb by 0.021 and 0.033, respectively. Indeed, Tasks 1 and 3 help solving it. Generalization in neural nets improved because the net learned to better represent underlying regularities of the domain. Extra outputs inject rule hints into networks about what they should learn. This is MTL where the extra tasks are carefully engineered to coerce the net to learn specific internal representations. The extra error terms constrain what is learned to satisfy desired properties of the other task. MTL provides a benefit with time series data because predictions at different time scales often partially depend on different processes. When learning a task with a short time scale, the learner may find it difficult to recognize the longer-term processes, and vice-versa. Training both scales on a single net improves the chances that both short- and long-term processes will be learned and combined to make predictions.

On another hand, the data resolution provided by the connected vehicles technology is another reason behind the high performance of our model . For instance, if the objective is to forecast traffic in 5-min periods into the future, then the best data resolution to be used is 5-min. Therefore, aggregation of high-resolution raw data into lower resolution levels is a common practice in short-term traffic forecasting studies. Many studies aggregate raw loop-detector data into 15-min, 10-min, and 5-min periods before the forecasting models are applied. In our case, since data are exchanged between connected vehicles every 0.1 seconds, we face the opposite case where we aggregate into high-resolution. We believe that in our case we don't loose accuracy as there is no extrapolation done.

Finally, we carefully analysed the instances where the MTL model made mistakes. We found that mistakes were made mostly from incident and workzone scenarios. In fact, two traffic situations can be represented by the same input features, but the output label of one of the

tasks can be different. For example, in traffic situation 1, the label at t+15 can be the same as the label t+15 in the traffic situation 2 but the label t+20 is different in both situations although they have the same input. In fact, we monitor the same network region at same time in the traffic situation 1 and if there is an incident on road segment 1 at the top of the street at time t1, we collect flows F1-F8 on all segments and establish the one-hot vector of S1 as being 1 0 0 0 0 0 and all others S2-S8 as being 000000. We construct the input feature vector of this situation. In another scenario at the same time t1, for the same network region, in traffic situation 2, there is an incident again on road segment 1 but this time at the bottom of the street. We collect flows F1-F8 on all segments. We find that they are the same as those in traffic scenario1 at the same time. We establish the one-hot vector of S1 as being again 1 0 0 0 0 0 and all others S2-S8 as being 000000. This results in the same input feature vector as previously. We have to label each feature vector with future flows. We see that flows at t1+15 for both scenarios are similar but flows for t1+20 are different. In this context, the incident happening on the street had the same effect on flow in a short period but later on it cleared better. The position of the incident on the road segment had an impact on flow later on in time because maybe meanwhile, the incident cleared better. We analysed the situation and found lots of vectors in the dataset that behave like this in the incident and workzone scenarios only. One way to solve this problem might be by changing the network architecture. In fact, the MTL net presented in this article use fully connected hidden layer shared equally by all tasks. Sometimes, more complex net architectures work better. For example, sometimes it is beneficial to have a small private hidden layer for the main task, and a larger hidden layer shared by both the main task and extra tasks. But too many private hidden layers (e.g., a private hidden layer for each task) reduce sharing and the benefits of MTL. We do not currently have principled ways to determine what architecture is best for each problem. Fortunately, simple architectures often work well, even if not optimally.

6.6 CONCLUSION

We proposed a Short-term Traffic flow Prediction (STP) framework for urban road networks so that traffic managers take early actions to control the flow and prevent the congestion state. The framework is semi-centralized because on one hand, connected vehicles collect and propagate data via the ad hoc networks they form between each other along a route. And on another hand, a road side unit (RSU) is installed on a target road segment and collects data for a period of time to get a clearer picture about the traffic on the road. To cope with the fact that current research on traffic prediction mainly focuses on data traffic history and neglects other conditions affecting traffic, in this paper, we showed how connected vehicles

technology allow advanced modelling by integrating the impact of the various events that CVs realistically encountered on segments along their trajectory into the forecasting of flow. We have studied how these critical events will affect the future traffic flow prediction and we solved the STP problem with a neural network. For a complex urban transportation system, one single hidden layer does not provide enough accuracy as it does not describe in detail the complicated relations between inputs and outputs. A deep architecture showed its advantage in dealing with these complicated relations. Supervised learning methods have been proposed in the literature for the task of STP however we showed how the design of our reactive approach by tackling the problem by learning the target Deep Neural Networks (DNN) in a multitask learning technique (MTL) improved prediction accuracy. In fact, the full detail of what is being learned for all tasks is available to all tasks because all tasks are being learned at the same time. Our shared semantic representation provides an important advantage over previous MTL applications, whose subtasks share a less consistent semantic representation. Our experiments on synthetic dataset show that the results of our approach significantly outperforms state-of-the-art ARIMA time series and baseline classifiers, with an average root-mean-square error (RMSE) of 0.05. Compared to single task learning with Artificial Neural Network (ANN), ANN had a lower performance (0.113 for RMSE) than MTL. In addition to improved performance, the proposed framework provides a flexible solution for general vehicular prediction tasks.

One limitation of the proposed machine learning prediction model include the fact that it is possible to increase the number of tasks in MTL and this hurts performance instead of helping it. MTL is a source of inductive bias. Some inductive biases help. Some inductive biases hurt. It depends on the problem. MTL is a tool that must be tested on each problem. Between mediocre performance on all tasks and optimal performance on one task, we optimized performance on one task, and allowed performance on the extra tasks to degrade. Also, real-world scenarios are much more irregular and noisier than synthetic scenarios. Scenarios used within the development of methods similar to the ones we developed could be publicly made available to show the benefits or limits of the new methods. It may also be noted that real-world scenarios are usually found to be more appealing and to convince a viewing person more than synthetic scenarios, especially in the transportation domain.The proposed methods in this article will require a 100% market penetration rate of the connected vehicles technology to obtain the full benefits of the connected vehicles technology and ensure satisfactory performance. This is actually more of a challenge to the success of our proposed solutions than a limitation.

As a future direction, to improve generalization performance, the framework could be easily extended to incorporate more features pertaining to the topology, contextual data about

seasonality such as day of week, whether a holiday is in progress. Also, in terms of security, at its current stage, the proposed framework represents a single point of failure as a faulty RSU means no prediction can be made on an entire target road segment. A prediction model should be able to give good predictions even in case of attacks of denial of service (DoS) on RSUs. Finally, the ultimate forecasting model takes into account the intention of travel of the individual as it is what is behind the behaviors of individual mobility. In fact, vehicles transit from one origin to a destination according to an intention of travel. A hybrid methodology that merges origin-destination matrices with combination of inter-vehicle communications feedback for data collection can better capture the real-time traffic state. Connected vehicles can capture the itinerary from each origin location to the intended destination in real time. The fast evolution of inter-vehicle communications, hence the potential of using connected vehicles with highly distributed algorithms are expected to provide near-optimal global solutions.

CHAPTER 7 GENERAL DISCUSSION

In this chapter, we first recall the research objectives that we declared at the beginning of the thesis and evaluate the extent to which they were achieved. Then, we take a critical look at the overall results of our work to better understand the scope and impacts of our research through the presented contributions. Finally, we discuss the limitations of our work.

7.1 Objectives achievement

The main objective of this thesis was to optimize the traffic flow in the transportation system in order to mitigate congestion. More specifically, the following objectives were attained :

— **Collect measurable traffic features extracted by an advanced monitoring technology capable of aggregating microscopic and macroscopic traffic variables at various levels of granularity.** In the first phase of our work in this thesis was emphasized how the duration, timing and location of non-recurrent congestion (NRC) in an urban network varies a lot making it difficult to monitor traffic in real time with conventional mechanisms. To this end, firstly, a thorough analysis of a set of unique traffic features representative of each type of NRC was undertaken. Specifically, incidents and workzones were essentially characterized by the presence of problematic spots on the traffic road segment on which the vehicle is travelling. The vehicle's variation in trajectory travel time, speed and gap was found to be representative of inclement weather. And special events were mainly characterised by their impact region and demand surge. Afterwards, in order to extract those traffic features, distributed monitoring was undertaken which refers to the process by which macroscopic and microscopic traffic variables were collected by the vehicles themselves. Connected vehicles were used as a next generation sensing technology, and allowed for distributed advanced monitoring. Due to real-time constraints, they were the information extraction technique needed to extract the transport-relevant parameters. This paved the way for the monitoring of trajectory data and enabled to scale to larger areas.

— **Propose classification models based on the traffic features collected for inference on the cause of congestion.** While existing methods only quantify the spatial and temporal impact of the detected NRC, we anticipated that understanding the cause of urban congestion is a prerequisite for deriving policies and management

plans so that appropriate proactive strategies can be set in place in order to return traffic state back to normality. To estimate the cause of congestion, we proposed a classification problem where each vehicle experiencing excessive congestion infers whether it is due to a weather condition, incident, workzone, recurrent or a special event. Because of the traffic features they collected via the connected vehicles technology, the vehicles in our framework were context aware and able to consider multiple adequate explanatory sources of information, particularly in dynamic urban environment. To solve the classification problem, machine learning models were designed and aimed at identifying the specific type of NRC based on the set of unique features experienced by a vehicle. The monitoring of traffic via connected vehicles was a cost-effective flexible solution and provided crucial help to increase the estimation accuracy of the classifiers we proposed. The classifiers were trained on synthetic data extended from the real case study of the Cologne scenario and their performance in terms of accuracy of classification demonstrated the robustness of our scheme.

— **Design an algorithm for the real-time assessment and evaluation of road traffic condition.** In the last part of the first phase of our thesis, we sought to further make use of the data collected by the vehicles to guarantee better monitoring of road traffic on heterogeneous networks. We enabled vehicles to not only detect excessive congestion and identify its cause but to also propagate the cause of NRC detected in its surrounding via connected vehicles. To this end, not only distributed advanced monitoring was employed but continuous advanced monitoring of the traffic condition along all roads of the traffic network was enabled since duration and timing of traffic events varies a lot. The purpose was to assess if the temporary induced traffic change related to an event can be mitigated in a short period or does the event represent a permanent change representing an NRC. The resulting algorithm highlighted the monitoring, aggregating, analysis and dissemination procedures done on board of each vehicle and proved that vehicles can evaluate collectively the cause of congestion experienced.

— **Simulate scenarios extended from a realistic urban city vehicular motion traces in order to build a synthetic dataset to feed the models for learning purposes.** In the second phase of our work, we stressed how economic issues, lack of large scale deployment and technology limitations made simulation the main choice in the validation of models based on vehicular ad hoc networks. Our work was the first of its kind to use both a microscopic urban mobility simulator, and a network simulator for the simulation of communication between CVs to generate a dataset

for learning. The real-world traffic of the Travel and Activity PAtterns Simulation (TAPAS) Cologne scenario [45], which is considered a 'complex network' that mimics the real-life context of vehicle mobility, was implemented in the microscopic simulator and was considered the base scenario. The mobility simulator needed two inputs : The Road Network of the city of Cologne is imported from the OpenStreetMap (OSM) database and the Traffic Demand which is the car trips of the base scenario. By calibrating the inputs of the mobility simulator, we were able to create extended scenarios mounted on top of the base scenario to model atypical traffic conditions such as weather, incident, workzone and special event. The output was the movement of vehicular nodes in a large urban network and data such as the acceleration, density, flows, gap between vehicles and other microscopic parameters at a vehicle level. From the simulation data collected by each vehicle, we extracted features constituting an instance of the train dataset called synthetic, in order to eventually feed the machine learning models.

— **Implement a cooperation process to increase estimation accuracy because traffic is multifaceted and to conceal the fact that individually, vehicles have partial knowledge about the road condition.** When congestion occurs, the vehicle tries to estimate the cause based on its experience. In the second phase of our work, we exposed the consequences of an assessment and classification done locally at a vehicle level. If one vehicle sends a false alarms, it spreads uncertainty among vehicles and this in turn causes more congestion. We tried to solve the problem of false alarms because the side effects of false alarms on the congestion level are a serious challenge and sending false information disrupts the proper network operation. We proposed methods to obtain deeper insight on the cause of traffic congestion using cooperation between CVs. Besides cooperation, we proposed that an evaluation process has to take place after data sensing and before data fusion and aggregation. We added this layer to address the vulnerability of fusion algorithms and also to lower the side effects of false alarms. The result was a voting procedure and the use of belief functions to increase accuracy of estimation. Finally, we explored the collected data for learning purposes by building a dataset and extracting relationships and knowledge via data mining techniques to elaborate a decision collectively. The performance evaluation showed that the mining technique enhanced estimation accuracy and detection time. It also decreased false alarms.

— **Propose a traffic flow prediction framework taking into account historical flows as well as innovative features, such as real-time reports from connec-**

ted vehicles and travel time along a trajectory for accurate forecasting of flow in an urban network. In the last phase of our work, we anticipated that knowing the flow of traffic heading toward a destination will give more insights about the expected demands in the near future. And this in return will allow traffic managers to take early actions to control the traffic flow and prevent the congestion state. Current models that address this traffic flow prediction problem needed to be improved so as to allow fast and more accurate prediction. In addition to the fact that they are based on data collected from a variety of traditional traffic sensors such as lidar, radar, and video from surveillance cameras. As the data originate from different sources, their conversion poses a problem. Besides, another obstacle is the amount of data collected which is increasing exponentially, and the complexity of the data. In sum, data aggregation poses many challenges when a variety of data sources are required in the process of data collection. We solved this traffic flow prediction problem by taking into account historical flows as well as innovative features, such as real-time reports from connected vehicles technology and travel time along a trajectory in the process of data collection. Also, another problem with current traffic flow prediction models is their inadaptability of detecting and tracking the traffic patterns changes. There is a new pattern every time a non recurrent congestion occurs in the traffic flow and in this case, the model is not able to predict as accurately as when there is recurrent congestion. Existing approaches do not adapt to the varying traffic situations because their distribution are memoryless, and they need a structure that will characterize the system at each step, not independently from the prior stage. We proposed a Deep Neural Networks (DNNs), and tackled the problem by learning the target DNN in a multitask learning technique. The model learns a representation that takes into account the various events that vehicles realistically encounter on the segments along their trajectory. The proposed model updates from its normal path and tracks the changed traffic pattern, generating forecasts according to the new traffic pattern.

7.2 Results analysis

A quick analysis of all the results obtained shows that they are more than satisfactory. One of the first contributions is the definition of a set of qualitative and quantitative features that describe the real-time traffic state experienced by any vehicle along its trajectory via the connected vehicles technology. In fact, when experiments were done anywhere on the road traffic network, the connected vehicles were able to detect the events and thus proved to be scalable. Also, they were able to report it quickly and this timely detection of events attested

the efficiency of this real time advanced monitoring technology.

This complex approach to traffic monitoring, although carried out by each vehicle by a very simple algorithm, lays the bases for a much more precise evaluation of the road traffic condition. In fact, when the traffic features extracted by the connected vehicles were fed to the machine learning methods we proposed for the inference on the cause of congestion, the results of the classifiers showed estimation accuracy ranging from 87.63% up to 89.51%. This contribution can assist transportation agencies in reducing urban congestion by developing effective congestion mitigation strategies knowing the root causes of congestion that are affecting their facilities.

However, since the best classifier had 89.51% of correctly classified instances, we carefully analysed the instances where the classifier made mistakes. We found a pattern and we sought to further improve the estimation accuracy. Another of our major contributions was the implementation of a cooperation process to increase estimation accuracy. The results showed that by adding an evaluation layer before fusion can take place on board of each vehicle not only increased accuracy but also lowered false alarms that are comparable to security threats on the traffic network. In fact, the distributed data mining techniques via connected vehicles to elaborate collectively a decision concerning the cause of traffic congestion on a road network improved the level of knowledge from exchanged messages and helped obtain deeper insight on traffic condition.

Finally, we make use of the contributions above, mainly, from the fact that vehicles can classify cooperatively the cause of any congestion encountered along their trajectory, to solve a traffic flow prediction task on any target road segment. The results of our proposed traffic flow prediction framework showed that taking into account historical flows as well as innovative features, such as real-time reports from connected vehicles and travel time along a trajectory for accurate forecasting of flow in an urban network improves prediction accuracy when compared to state-of-the-art time series approach or baseline classifiers. This is due to the fact that tasks in our proposed model benefit each other mutually, something a linear sequence cannot capture and thus ignores a potentially rich source of information available, the information contained in the training signals of other tasks drawn from the same domain. If the tasks can share what they learn, the learner may find it is easier to learn them together than in isolation. In fact, we found that generalization in artificial neural nets improved because the nets learned to better represent underlying regularities of the road traffic network. Extra outputs inject rule hints into the model about what they should learn. This is MTL where the extra tasks are carefully engineered to coerce the net to learn specific internal representations.

7.3 Limitations

The proposed methods in this thesis will require a market penetration rate of the connected vehicles technology between 63% and 75% to obtain the full benefits of V2V communications and ensure satisfactory performance. A low penetration rate implies that few vehicles are equipped with transceivers and this leads to network fragmentation in the context of ad hoc networks. As a consequence, vehicles have to wait to be in the communication range of each other for the methods to be effective. The performance strongly depends on the penetration rate of participating vehicles. This is actually more of a challenge to the success of our proposed solutions than a limitation.

Also, real-world scenarios are much more irregular and noisier than synthetic scenarios. They may also include some peculiarities specific for the area, often dictated by a chosen long-term traffic management strategy. Also, scenarios used within the development of methods similar to the ones we developed could be publicly made available to show the benefits or limits of the new methods. It may also be noted that real-world scenarios are usually found to be more appealing and to convince a viewing person more than synthetic scenarios, especially in the transportation domain.

Finally, limitations of the proposed machine learning prediction model include :

— It yields worst-case bounds that are too loose to insure extra tasks will help. For example, it is possible to increase the number of tasks and this hurts performance instead of helping it. MTL does not always improve performance. MTL is a source of inductive bias. Some inductive biases help. Some inductive biases hurt. It depends on the problem. MTL is a tool that must be tested on each problem.

— Where tradeoffs can be made between mediocre performance on all tasks and optimal performance on any one task, usually it is best to optimize performance on tasks one at a time, and allow performance on the extra tasks to degrade.

— It still is a blackblox approach, it lacks a well-defined notion of task relatedness. Also, we usually find that optimal performance requires increasing the number of units in the shared hidden layer as the number of tasks increases. This conflicts with assumptions made by the theory that the hidden layer size remain constant as the number of tasks increases.

CHAPTER 8 CONCLUSION

With the increasing number of vehicles and limited road network expansion, the urban traffic congestion is growing at an alarming rate. In this thesis, we proposed to optimize the traffic flow in the transportation system in order to mitigate congestion. We proceeded to a road traffic congestion analysis via connected vehicles. The main originality of this thesis lies in the use of the next generation sensing technology of connected vehicles to identify road traffic events on the basis of exchanging traffic flow data between vehicles. This novel approach in this domain allowed the real-time distributed detection and classification of the components of congestion in urban traffic. Moreover, if connected vehicles can detect congestion and cooperatively attribute a possible cause to it, we showed that they can transfer this knowledge in real time to an entity able to accurately predict flow on a road segment. The traffic flow prediction framework we introduced aimed at evaluating anticipated traffic flow at future time frames on a target road segment based on real time feeds provided by connected vehicles and historical data. Traffic flow prediction allows advanced modelling because knowing the volume of traffic heading toward a destination will give more insights about the expected demands in the near future. The proposed models and framework will help infrastructure authorities improve the network traffic flow and thus reduce traffic congestion.

As a future direction, more features pertaining to the topology, contextual data about seasonality such as day of week, whether a holiday is in progress, can be incorporated into the models and can be informative about traffic flows in order to improve accuracy. Moreover, the ultimate forecasting model takes into account the intention of travel of the individual as it is what is behind the behaviours of individual mobility. Recently, a conceptual framework of artificial transportation systems focused on behaviour or intention of travel, to model and simulate artificial societies on a participative basis [99]. This framework can be used to implement a hybrid methodology that merges inter-vehicle communications with the intention of travellers to enhance the estimation of road traffic and to increase the accuracy of information retrieval. The main idea behind the framework is to integrate various transportation models into "artificial" transportation systems and convert computers into experimental "fields" for transportation analysis and decision making and evaluations. The keys to successful applications of such artificial transportations are the availability of agent-based programming and modeling, large scale distributed computing techniques, and new concepts and methods developed in complex systems, such as, artificial societies, computational experiments, and parallel systems.

REFERENCES

[1] S. Djahel, R. Doolan, G.-M. Muntean, and J. Murphy, "A communications-oriented perspective on traffic management systems for smart cities : challenges and innovative approaches," *Communications Surveys & Tutorials, IEEE*, vol. 17, no. 1, pp. 125–151, 2015.

[2] T. V. Mathew, "Fundamental relations of traffic flow."

[3] L. A. Klein, M. K. Mills, D. Gibson, and L. A. Klein, "Traffic detector handbook : Volume ii," United States. Federal Highway Administration, Tech. Rep., 2006.

[4] S. Djahel, R. Doolan, G. M. Muntean, and J. Murphy, "A communications-oriented perspective on traffic management systems for smart cities : Challenges and innovative approaches," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 125–151, Firstquarter 2015.

[5] A. H. Chow, A. Santacreu, I. Tsapakis, G. Tanasaranond, and T. Cheng, "Empirical assessment of urban traffic congestion," *Journal of advanced transportation*, vol. 48, no. 8, pp. 1000–1016, 2014.

[6] J. Asamer, H. J. van Zuylen, and B. Heilmann, "Calibrating car-following parameters for snowy road conditions in the microscopic traffic simulator vissim," *Intelligent Transport Systems, IET*, vol. 7, no. 1, pp. 114–121, 2013.

[7] M. Brackstone and M. McDonald, "Car-following : a historical review," *Transportation Research Part F : Traffic Psychology and Behaviour*, vol. 2, no. 4, pp. 181–196, 1999.

[8] S. P. Latoski, W. M. Dunn Jr, B. Wagenblast, J. Randall, and M. D. Walker, "Managing travel for planned special events," Tech. Rep., 2003.

[9] S. Kwoczek, S. Di Martino, and W. Nejdl, "Predicting and visualizing traffic congestion in the presence of planned special events," *Journal of Visual Languages & Computing*, vol. 25, no. 6, pp. 973–980, 2014.

[10] B. Anbaroğlu, T. Cheng, and B. Heydecker, "Non-recurrent traffic congestion detection on heterogeneous urban road networks," *Transportmetrica A : Transport Science*, vol. 11, no. 9, pp. 754–771, 2015.

[11] H. Kamal, M. Picone, and M. Amoretti, "A survey and taxonomy of urban traffic management : Towards vehicular networks," *arXiv preprint arXiv :1409.4388*, 2014.

[12] M. Gramaglia, O. Trullols-Cruces, D. Naboulsi, M. Fiore, and M. Calderon, "Vehicular networks on two madrid highways," in *Sensing, Communication, and Networking (SECON), 2014 Eleventh Annual IEEE International Conference on*, 2014, pp. 423–431.

[13] L. Li, X. Su, Y. Wang, Y. Lin, Z. Li, and Y. Li, "Robust causal dependence mining in big data network and its application to traffic flow predictions," *Transportation Research Part C : Emerging Technologies*, vol. 58, pp. 292–307, 2015.

[14] T. Darwish and K. A. Bakar, "Traffic density estimation in vehicular ad hoc networks : A review," *Ad Hoc Networks*, vol. 24, pp. 337–351, 2015.

[15] Y. Gu, L.-T. Hsu, and S. Kamijo, "Towards lane-level traffic monitoring in urban environment using precise probe vehicle data derived from three-dimensional map aided differential gnss," *IATSS Research*, 2018.

[16] X. Yang, Z. Fang, L. Yin, J. Li, Y. Zhou, and S. Lu, "Understanding the spatial structure of urban commuting using mobile phone location data : A case study of shenzhen, china," *Sustainability*, vol. 10, no. 5, p. 1435, 2018.

[17] G. S. Khekare and A. V. Sakhare, "A smart city framework for intelligent traffic system using vanet," in *Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013 International Multi-Conference on*, 2013, pp. 302–305.

[18] Q. Yuan, Z. Liu, J. Li, J. Zhang, and F. Yang, "A traffic congestion detection and information dissemination scheme for urban expressways using vehicular networks," *Transportation Research Part C : Emerging Technologies*, vol. 47, pp. 114–127, 2014.

[19] M. B. Younes and A. Boukerche, "A performance evaluation of an efficient traffic congestion detection protocol (ecode) for intelligent transportation systems," *Ad Hoc Networks*, vol. 24, pp. 317–336, 2015.

[20] C. Wang and H.-M. Tsai, "Detecting urban traffic congestion with single vehicle," in *Connected Vehicles and Expo (ICCVE), 2013 International Conference on*, 2013, pp. 233–240.

[21] L. Zhang, D. Gao, W. Zhao, and H.-C. Chao, "A multilevel information fusion approach for road congestion detection in vanets," *Mathematical and Computer Modelling*, vol. 58, no. 5, pp. 1206–1221, 2013.

[22] S. A. Vaqar and O. Basir, "Traffic pattern detection in a partially deployed vehicular ad hoc network of vehicles," *Wireless Communications, IEEE*, vol. 16, no. 6, pp. 40–46, 2009.

[23] E. J. Horvitz, J. Apacible, R. Sarin, and L. Liao, "Prediction, expectation, and surprise : Methods, designs, and study of a deployed traffic forecasting service," *arXiv preprint arXiv :1207.1352*, 2012.

[24] R. Bauza and J. Gozálvez, "Traffic congestion detection in large-scale scenarios using vehicle-to-vehicle communications," *Journal of Network and Computer Applications*, vol. 36, no. 5, pp. 1295–1307, 2013.

[25] J. Kwon, T. Barkley, R. Hranac, K. Petty, and N. Compin, "Decomposition of travel time reliability into various sources : incidents, weather, work zones, special events, and base capacity," *Transportation Research Record : Journal of the Transportation Research Board*, no. 2229, pp. 28–33, 2011.

[26] Y. A. Al-Khassawneh and N. Salim, "On the use of data mining techniques in vehicular ad hoc network," in *Advanced Machine Learning Technologies and Applications*. Springer, 2012, pp. 449–462.

[27] M. Miller and C. Gupta, "Mining traffic incidents to forecast impact," in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, 2012, pp. 33–40.

[28] R. Hranac, E. Sterzin, D. Krechmer, H. A. Rakha, M. Farzaneh, M. Arafeh *et al.*, "Empirical studies on traffic flow in inclement weather," 2006.

[29] Y. Hou, P. Edara, and C. Sun, "Traffic flow forecasting for urban work zones," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 16, no. 4, pp. 1761–1770, 2015.

[30] J. W. Wedel, B. Schünemann, and I. Radusch, "V2x-based traffic congestion recognition and avoidance," in *2009 10th International Symposium on Pervasive Systems, Algorithms, and Networks*, 2009, pp. 637–641.

[31] S. Manvi, M. Kakkasageri, and J. Pitt, "Multiagent based information dissemination in vehicular ad hoc networks," *Mobile Information Systems*, vol. 5, no. 4, pp. 363–389, 2009.

[32] F. Terroso-Saenz, M. Valdes-Vela, C. Sotomayor-Martinez, R. Toledo-Moreo, and A. F. Gomez-Skarmeta, "A cooperative approach to traffic congestion detection with complex event processing and vanet," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 13, no. 2, pp. 914–929, 2012.

[33] F. C. Pereira, A. L. Bazzan, and M. Ben-Akiva, "The role of context in transport prediction," *IEEE Intelligent Systems*, no. 1, pp. 76–80, 2014.

[34] E. A. Gamati, E. Peytchev, and R. Germon, "Traffic condition detection algorithm (tcda) for vanet nodes in wireless intelligent transportation information systems." in *ECMS*, 2011, pp. 459–465.

[35] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing, "Discovering spatio-temporal causal interactions in traffic data streams," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1010–1018.

[36] S. Thajchayapong, E. S. Garcia-Trevino, and J. A. Barria, "Distributed classification of traffic anomalies using microscopic traffic variables," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 14, no. 1, pp. 448–458, 2013.

[37] M. Abuelela and S. Olariu, "Automatic incident detection in vanets : a bayesian approach," in *Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th*, 2009, pp. 1–5.

[38] H. Qin, X. Lu, Y. Wang, G. Wang, W. Zhang, and Y. Zhang, "Heterogeneity-aware design for automatic detection of problematic road conditions," in *Mobile Adhoc and Sensor Systems (MASS), 2011 IEEE 8th International Conference on*, 2011, pp. 252–261.

[39] V. Cherfaoui, T. Denoeux, and Z. L. Cherfi, "Distributed data fusion : application to confidence management in vehicular networks," in *2008 11th International Conference on Information Fusion*, June 2008, pp. 1–8.

[40] M. B. Farah, D. Mercier, E. Lefevre, and F. Delmotte, "Towards a robust exchange of imperfect information in inter-vehicle ad-hoc networks using belief functions," in *2011 IEEE Intelligent Vehicles Symposium (IV)*, June 2011, pp. 436–441.

[41] ——, "Exchanging dynamic and imprecise information in v2v networks with belief functions," in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, Oct 2013, pp. 967–972.

[42] M. B. Farah, D. Mercier, F. Delmotte, and Éric Lefèvre, "Methods using belief functions to manage imperfect information concerning events on the road in vanets," *Transportation Research Part C : Emerging Technologies*, vol. 67, pp. 299 – 320, 2016. [Online]. Available : http://www.sciencedirect.com/science/article/pii/S0968090X1600067X

[43] M. H. Arbabi and M. Weigle, "Using vehicular networks to collect common traffic data," in *Proceedings of the Sixth ACM International Workshop on VehiculAr InterNETworking*, ser. VANET '09. New York, NY, USA : ACM, 2009, pp. 117–118. [Online]. Available : http://doi.acm.org/10.1145/1614269.1614289

[44] M. Piórkowski, M. Raya, A. L. Lugo, P. Papadimitratos, M. Grossglauser, and J.-P. Hubaux, "Trans : Realistic joint traffic and network simulator for vanets," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 12, no. 1, pp. 31–33, Jan. 2008. [Online]. Available : http://doi.acm.org/10.1145/1374512.1374522

[45] S. Uppoor and M. Fiore, "Large-scale urban vehicular mobility for networking research," in *Vehicular Networking Conference (VNC), 2011 IEEE*, 2011, pp. 62–69.

[46] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal arima process : Theoretical basis and empirical results," *Journal of transportation engineering*, vol. 129, no. 6, pp. 664–672, 2003.

[47] T. Pan, A. Sumalee, R.-X. Zhong, and N. Indra-Payoong, "Short-term traffic state prediction based on temporal–spatial correlation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1242–1254, 2013.

[48] M. M. Hamed, H. R. Al-Masaeid, and Z. M. B. Said, "Short-term prediction of traffic volume in urban arterials," *Journal of Transportation Engineering*, vol. 121, no. 3, pp. 249–254, 1995.

[49] M. Van Der Voort, M. Dougherty, and S. Watson, "Combining kohonen maps with arima time series models to forecast traffic flow," *Transportation Research Part C : Emerging Technologies*, vol. 4, no. 5, pp. 307–318, 1996.

[50] C. Kongcharoen and T. Kruangpradit, "Autoregressive integrated moving average with explanatory variable (arimax) model for thailand export," in *33rd International Symposium on Forecasting, South Korea*, 2013, pp. 1–8.

[51] J. D. Hamilton, *Time series analysis*. Princeton university press Princeton, 1994, vol. 2.

[52] F.-M. Tseng and G.-H. Tzeng, "A fuzzy seasonal arima model for forecasting," *Fuzzy Sets and Systems*, vol. 126, no. 3, pp. 367–376, 2002.

[53] B. L. Smith and M. J. Demetsky, "Short-term traffic flow prediction models-a comparison of neural network and nonparametric regression approaches," in *Systems, Man, and Cybernetics, 1994. Humans, Information and Technology., 1994 IEEE International Conference on*, vol. 2, 1994, pp. 1706–1709.

[54] T. Wu, K. Xie, D. Xinpin, and G. Song, "A online boosting approach for traffic flow forecasting under abnormal conditions," in *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on*, 2012, pp. 2555–2559.

[55] M. Castro-Neto, Y.-S. Jeong, M.-K. Jeong, and L. D. Han, "Online-svr for short-term traffic flow prediction under typical and atypical traffic conditions," *Expert systems with applications*, vol. 36, no. 3, pp. 6164–6173, 2009.

[56] C. Ledoux, "An urban traffic flow model integrating neural networks," *Transportation Research Part C : Emerging Technologies*, vol. 5, no. 5, pp. 287–300, 1997.

[57] J.-l. Zhang and X.-y. Wang, "Traffic flow prediction method based on non-linear hybrid model," *Computer Engineering*, vol. 5, p. 075, 2010.

[58] C. Ulbricht, "Multi-recurrent networks for traffic forecasting," in *AAAI*, 1994, pp. 883–888.

[59] M. G. Karlaftis and E. I. Vlahogianni, "Statistical methods versus neural networks in transportation research : Differences, similarities and some insights," *Transportation Research Part C : Emerging Technologies*, vol. 19, no. 3, pp. 387–399, 2011.

[60] H. Chen and S. Grant-Muller, "Use of sequential learning for short-term traffic flow forecasting," *Transportation Research Part C : Emerging Technologies*, vol. 9, no. 5, pp. 319–336, 2001.

[61] Y. Chen, L. Shu, and L. Wang, "Traffic flow prediction with big data : A deep learning based time series model," in *Computer Communications Workshops (INFOCOM WKSHPS), 2017 IEEE Conference on*, 2017, pp. 1010–1011.

[62] W. Huang, G. Song, H. Hong, and K. Xie, "Deep architecture for traffic flow prediction : deep belief networks with multitask learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 2191–2201, 2014.

[63] S. Sun, R. Huang, and Y. Gao, "Network-scale traffic modeling and forecasting with graphical lasso and neural networks," *Journal of Transportation Engineering*, vol. 138, no. 11, pp. 1358–1367, 2012.

[64] S. Sun, "Traffic flow forecasting based on multitask ensemble learning," in *Proceedings of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation*, 2009, pp. 961–964.

[65] F. Jin and S. Sun, "Neural network multitask learning for traffic flow forecasting," in *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, 2008, pp. 1897–1901.

[66] 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.

[67] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Optimized and meta-optimized neural networks for short-term traffic flow prediction : A genetic approach," *Transportation Research Part C : Emerging Technologies*, vol. 13, no. 3, pp. 211–234, 2005.

[68] Z. Wang, M. Lu, X. Yuan, J. Zhang, and H. Van De Wetering, "Visual traffic jam analysis based on trajectory data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2159–2168, 2013.

[69] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection : Techniques, systems and challenges," *computers & security*, vol. 28, no. 1-2, pp. 18–28, 2009.

[70] P. Poonia, V. Jain, and A. Kumar, "Short term traffic flow prediction methodologies : A review," *Mody University International Journal of Computing and Engineering Research*, vol. 2, no. 1, pp. 37–39, 2018.

[71] Y. Kamarianakis and P. Prastacos, "Space–time modeling of traffic flow," *Computers & Geosciences*, vol. 31, no. 2, pp. 119–133, 2005.

[72] D. Krajzewicz, G. Hertkorn, C. Rössel, and P. Wagner, "Sumo (simulation of urban mobility)-an open-source traffic simulation," in *Proceedings of the 4th Middle East Symposium on Simulation and Modelling (MESM20002)*, 2002, pp. 183–187.

[73] (Accessed : 2015-12-01) Speed Concepts : Informational Guide . http://safety.fhwa.dot.gov/speedmgt/ref_mats/fhwasa10001/fhwasa10001.pdf.

[74] K. P. Murphy, *Machine Learning : A Probabilistic Perspective*. The MIT Press, 2012.

[75] F. L. Hall, "Traffic stream characteristics," *Traffic Flow Theory. US Federal Highway Administration*, 1996.

[76] M. H. Rehmani and Y. Saleem, "Network simulator ns-2."

[77] "eWorld homepage," Accessed : 2016-04-01, http://eworld.sourceforge.net/.

[78] R. Caruana, "Multitask learning," in *Learning to learn*. Springer, 1998, pp. 95–133.

[79] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software : An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available : http://doi.acm.org/10.1145/1656274.1656278

[80] (Accessed : 2016-01-01) Mobility transformation center @ONLINE. http://www.mtc.umich.edu/deployments.

[81] W. S. Manjoro, M. Dhakar, and B. K. Chaurasia, "Traffic congestion detection using data mining in vanet," in *2016 IEEE Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, March 2016, pp. 1–6.

[82] M. Gramaglia, M. Calderon, and C. Bernardos, "Abeona monitored traffic : Vanet-assisted cooperative traffic congestion forecasting," *Vehicular Technology Magazine, IEEE*, vol. 9, no. 2, pp. 50–57, 2014.

[83] R. A. Mallah, A. Quintero, and B. Farooq, "Distributed classification of urban congestion using vanet," *IEEE Transactions on Intelligent Transportation Systems*, vol. PP, no. 99, pp. 1–8, 2017.

[84] M. S. Kakkasageri, S. S. Manvi, and J. Pitt, "Cognitive agent based critical information gathering and dissemination in vehicular ad hoc networks," *Wireless Personal Communications*, vol. 69, no. 4, pp. 1107–1129, Apr 2013. [Online]. Available : https://doi.org/10.1007/s11277-012-0623-5

[85] Y. Dieudonné, B. Ducourthial, and S. M. Senouci, "Col : A data collection protocol for vanet," in *2012 IEEE Intelligent Vehicles Symposium*, June 2012, pp. 711–716.

[86] F. Sun, A. Dubey, and J. White, "Dxnat x2014; deep neural networks for explaining non-recurring traffic congestion," in *2017 IEEE International Conference on Big Data (Big Data)*, Dec 2017, pp. 2141–2150.

[87] B. Pan, U. Demiryurek, and C. Shahabi, "Utilizing real-world transportation data for accurate traffic prediction," in *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, 2012, pp. 595–604.

[88] J. Xu, D. Deng, U. Demiryurek, C. Shahabi, and M. Van Der Schaar, "Context-aware online spatiotemporal traffic prediction," in *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*, 2014, pp. 43–46.

[89] 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.

[90] J. Xu, D. Deng, U. Demiryurek, C. Shahabi, and M. Van Der Schaar, "Mining the situation : Spatiotemporal traffic prediction with big data," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 4, pp. 702–715, 2015.

[91] K. Y. Chan, T. S. Dillon, and E. Chang, "An intelligent particle swarm optimization for short-term traffic flow forecasting using on-road sensor systems," *IEEE Transactions on Industrial Electronics*, vol. 60, no. 10, pp. 4714–4725, 2013.

[92] C.-S. Li and M.-C. Chen, "Identifying important variables for predicting travel time of freeway with non-recurrent congestion with neural networks," *Neural Computing and Applications*, vol. 23, no. 6, pp. 1611–1629, 2013.

[93] N. Lu, N. Cheng, N. Zhang, X. Shen, and J. W. Mark, "Connected vehicles : Solutions and challenges," *IEEE internet of things journal*, vol. 1, no. 4, pp. 289–299, 2014.

[94] G. Dimitrakopoulos and P. Demestichas, "Intelligent transportation systems," *IEEE Vehicular Technology Magazine*, vol. 5, no. 1, pp. 77–84, 2010.

[95] L. Zhang, L. Jia, and W. Zhu, "Overview of traffic flow hybrid ann forecasting algorithm study," in *Computer Application and System Modeling (ICCASM), 2010 International Conference on*, vol. 1, 2010, pp. V1–615.

[96] H. C. Manual, "Highway capacity manual," *Washington, DC*, p. 11, 2000.

[97] S. Dunne and B. Ghosh, "Weather adaptive traffic prediction using neurowavelet models," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 370–379, 2013.

[98] N. S. Nafi, R. H. Khan, J. Y. Khan, and M. Gregory, "A predictive road traffic management system based on vehicular ad-hoc network," in *Telecommunication Networks and Applications Conference (ATNAC), 2014 Australasian*, 2014, pp. 135–140.

[99] F.-Y. Wang and S.-m. Tang, "Concepts and frameworks of artificial transportation systems," *Complex Systems and Complexity Science*, vol. 1, no. 2, pp. 52–59, 2004.